DESCUBRIMIENTO DE FACTORES ASOCIADOS AL DESEMPEÑO EN LAS PRUEBAS SABER 5 CON TÉCNICAS DESCRIPTIVAS DE MINERÍA DE DATOS

LEIDY MARCELA GÓMEZ MELO

YAZMÍN ALEXANDRA JARAMILLO CÓRDOBA

UNIVERSIDAD TECNOLÓGICA DE PEREIRA-CONVENIO UNIVERSIDAD DE NARIÑO

FACULTAD DE INGENIERÍA INDUSTRIAL

MAESTRÍA EN INVESTIGACIÓN DE OPERACIONES Y ESTADÍSTICA

PASTO

2017

DESCUBRIMIENTO DE FACTORES ASOCIADOS AL DESEMPEÑO EN LAS PRUEBAS SABER 5 CON TÉCNICAS DESCRIPTIVAS DE MINERÍA DE DATOS

LEIDY MARCELA GÓMEZ MELO

YAZMÍN ALEXANDRA JARAMILLO CÓRDOBA

TRABAJO DE GRADO PRESENTADO COMO REQUISITO PARA OPTAR EL TITULO DE MAGISTER EN INVESTIGACIÓN DE OPERACIONES Y ESTADÍSTICA

DIRECTOR:

PhD. RICARDO TIMARÁN PEREIRA

UNIVERSIDAD TECNOLÓGICA DE PEREIRA-CONVENIO UNIVERSIDAD DE NARIÑO

FACULTAD DE INGENIERÍA INDUSTRIAL

MAESTRÍA EN INVESTIGACIÓN DE OPERACIONES Y ESTADÍSTICA

PASTO

2017

Agradecimientos

Al PhD. Ricardo Timarán Pereira por su arduo trabajo de asesoría, sus valiosos consejos y recomendaciones.

Al MSc. Hernán García por ser el gestor y parte importante del proceso de formación.

Al PhD. José Soto Mejía por darnos la oportunidad de la formación en el campo de la Estadística e Investigación de Operaciones.

Al Ing. Wilson Gómez por su valiosa colaboración y apoyo en las tareas de programación y procesamiento de datos.

A la Universidad Tecnológica de Pereira, la Universidad de Nariño y a todos los profesionales que nos brindaron sus conocimientos y orientaciones en pro del crecimiento intelectual, profesional y personal.

DEDICATORIA

Padre celestial, quiero dedicarte este trabajo de grado como símbolo de agradecimiento profundo por tu infinito amor, por tu bondad y misericordia; porque me llenaste de valor y fortaleza para emprender este reto y para que lo pueda culminar con éxito. Gracias, porque cada vez que quería rendirme estuviste presente manifestándome tu amor por medio de las personas que están a mi lado, regalándome una palabra de aliento o simplemente con un abrazo me demostraban su apoyo incondicional.

A mi padre Héctor, mi ángel que desde el cielo sigue cuidándome como la niña de sus ojos, ese ser que me enseñó a emprender grandes retos, y sé que sin importar las dificultades que tenga, él siempre estará para levantarme.

A mi madre Olga, una mujer luchadora, ejemplo de fortaleza ante las circunstancias que le trae la vida. Gracias por enseñarme que la vida es de esfuerzos para conseguir lo que te propones.

A mis hermanas Lilia, Marcela y Fernanda, quienes miran en mí un ejemplo a seguir. Sin saber que son ellas las que me demuestran ejemplos de superación.

A mi amado esposo Gustavo, que con su cariño, compañía y confianza permitió que logremos juntos la meta de crecer como familia, como persona y como profesional.

A mis hijos, Manuel y Julián, mi razón de ser, el motor de mi vida. Son ellos los que me impulsan a levantarme cada mañana, en cada caída, ante cualquier adversidad… ¡Gracias por ese amor puro y verdadero!

Yazmín Alexandra Jaramillo Córdoba

Dedicatoria

Al llegar a la culminación de mi maestría sin duda alguna el primer pensamiento de agradecimiento es a mi Padre Dios, porque sin él mi vida no tendría el rumbo que ha tomado, gracias Dios mío por bendecirme con tantas oportunidades de salir adelante formándome en el ámbito espiritual, personal y profesional.

Dedico esta meta cumplida a mis queridos padres Marcos e Ignacia, quienes con su ejemplo de vida y disciplina han sembrado en mí el deseo por emprender grandes metas y logre culminarlas con éxito.

A mis hermanos porque con sus palabras de aliento siempre me animan en los momentos de dificultad y quebranto.

A mi amor verdadero, mi Mune de quien he aprendido sobre el verdadero valor de la vida, el amor, la paciencia, la disciplina y la entrega desinteresada, quien ha visto en mí una mujer valiosa y quien me dio el valor para trabajar arduamente por cumplir mis metas. Tú y tu amor siempre serán la mayor inspiración de mi vida. Te dedico con todo el amor de mi alma este y cada uno de los triunfos que de seguro vendrán en un futuro no muy lejano.

Marcela Gómez

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Resumen

El objetivo de esta investigación fue descubrir factores asociados al desempeño académico en las competencias genéricas de las pruebas Saber 5° de los estudiantes de Instituciones Educativas de Colombia que presentaron estas pruebas en el periodo 2014 al 2016, utilizando técnicas descriptivas de minería de datos. Se utilizaron los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del Instituto Colombiano para la Evaluación de la Educación (ICFES). Se aplicó la metodología CRISP-DM, uno de los modelos más utilizado en los ambientes académico e industrial y la guía de referencia más ampliamente aplicada en el desarrollo de este tipo de proyectos. A partir del desarrollo de las fases de esta metodología, se obtuvo en primer lugar un repositorio de datos limpio y transformado, para las competencias de Lenguaje, Matemáticas, Ciencias Naturales y Competencias Ciudadanas. Se utilizaron las técnicas descriptivas de minería de datos Reglas de Asociación con el algoritmo Apriori y la técnica de Agrupamiento o *Clustering* con el algoritmo k-means, para descubrir factores asociados al desempeño académico. La gran mayoría de factores están asociados al desempeño mínimo en las competencias evaluadas. El conocimiento descubierto se incorporará al existente y se podrá integrar a los procesos de toma de decisiones del ICFES y de las instituciones gubernamentales y académicas que velan por la calidad de la educación en el País.

Palabras clave: calidad educativa, competencias genéricas, desempeño académico, minería de datos, metodología CRISP-DM, clustering, asociación, factores asociados, pruebas Saber 5.

Abstract

The goal of this research was to discover factors associated with academic performance in the generic skills of the Saber 5° tests of the students of Educational Institutions of Colombia who presented these tests in the period between 2014 and 2016, using descriptive data mining techniques. The socioeconomic, academic and institutional data stored in the databases of the Colombian Institute for the Evaluation of Education (ICFES) were used. The CRISP-DM methodology was applied. It is one of the most used models in the academic and industrial environments and the reference guide most widely applied in the development of this type of projects. From the development of the phases of this methodology, a clean and transformed data repository was first obtained for the skills of Language, Mathematics, Natural Sciences and Citizen Competencies. It used descriptive data mining techniques Association Rules with the Apriori algorithm and the Clustering technique with the k-means algorithm, to discover factors associated with academic performance. The vast majority of factors are associated with the minimum performance of the evaluated skills. The knowledge discovered will be incorporated into the existing one and can be integrated into the decision-making processes of ICFES and the governmental and academic institutions that ensure the quality of education in the country.

Key words: Quality of education, generic skills, academic performance, data mining, CRISP-DM methodology, clustering, association, associated factors, Saber 5 tests.

1. Aspectos Preliminares

1.1 Introducción

En la actualidad se han realizado algunos estudios teniendo en cuenta las Pruebas Saber, como el realizado por Chica, S., Galvis, D. y Ramírez, A. (2010), Torres, J., Pachajoa, L. y Pantoja, R. (2014) que fueron fundamentados en información procesada mediante técnicas estadísticas, donde esencialmente se consideran variables y relaciones primarias, sin tener en cuenta las verdaderas interrelaciones, que por lo general están ocultas y que únicamente se pueden descubrir utilizando un tratamiento más complejo de los datos, lo cual es posible con la Minería de Datos.

En esta investigación se propuso descubrir factores asociados al desempeño académico de los estudiantes de las instituciones educativas del país, que cursando grado quinto, presentaron las Pruebas Saber 5° en los años 2014 a 2016, a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES, con técnicas descriptivas de Minería de Datos.

La metodología a utilizar es CRISP-DM, la guía más ampliamente empleada en el desarrollo de proyectos de Minería de Datos, que contempla seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. El conocimiento descubierto se incorporará al existente y se podrá integrar a los procesos de toma de decisiones de las instituciones gubernamentales y educativas que velan por la calidad de la educación en la República de Colombia.

1.2 Planteamiento Del Problema

Una de las grandes problemáticas en Colombia desde tiempos atrás hasta la actualidad es la calidad de la educación; puesto que las condiciones familiares, sociales, económicas y culturales, que existe para la población estudiantil han influido de manera significativa en el

desempeño académico, Rodríguez, E. (2014), lo cual desencadena una baja calidad educativa reflejada específicamente en los resultados de las Pruebas Saber, con desempeños bajos e insuficientes, Jafet, C. & Martínez, C. (2016).

El Instituto Colombiano para la Evaluación de la Educación (ICFES) dio apertura a investigaciones que permitieron identificar variables relacionadas con el rendimiento de las pruebas, Como el informe realizado sobre Factores asociados en las Pruebas Saber de 5° y 9° (ICFES, 2011), en el cual se aplicaron técnicas estadísticas que permitieron visualizar elementos que inciden en el desempeño académico; para extender sus procesos de evaluación el ICFES dio paso al estudio de los factores asociados al rendimiento escolar utilizando modelos teóricos para explicar las relaciones existentes entre los elementos que determinan el aprendizaje, los cuales están presentes en tres niveles: instituciones educativas, aulas de clase y estudiantes (ICFES, 2016).

En Colombia se han realizado varios estudios entorno a las Pruebas Saber 5 como por ejemplo el realizado por Torres, J., Pachajoa, L. y Pantoja, R. (2014), Martín, S. (2015), Gutiérrez, Y. (2015) que buscan identificar las variables asociadas al rendimiento académico y en especial al desempeño de las Pruebas Saber 5; ellos asocian el resultado de las pruebas con la calidad educativa y afirman que tiene que ver no solamente con el estudiante si no también con el profesor. Para sus estudios tomaron como base únicamente una de las áreas fundamentales que para este caso fue ciencias naturales, matemáticas o Lenguaje respectivamente.

En otro estudio (ICFES, 2011), se analizaron los factores asociados de las pruebas de grado 5° y 9° en el cual una de las conclusiones o hallazgos fue que a mayor nivel socioeconómico de los alumnos y sus familias, mayor es el desempeño esperado en ambas áreas y grados evaluados en SABER 5° y 9°. Además, los estudiantes matriculados en colegios privados tienden a obtener

puntajes más altos en las pruebas, y las diferencias frente a quienes asisten a planteles oficiales se incrementan en la medida en que mejoran las condiciones socioeconómicas.

En esta investigación se pretende descubrir los factores asociados al rendimiento académico de los estudiantes de instituciones educativas del país que presentaron las Pruebas Saber 5° en el periodo 2014 a 2016 a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES, aplicando técnicas de Minería de Datos que permiten generar conocimiento encaminado a soportar las decisiones institucionales y gubernamentales para el mejoramiento de la calidad de la educación básica y media.

La presente investigación busca responder a la pregunta: ¿Cuáles son los factores asociados al desempeño académico de los estudiantes a nivel nacional que presentan las Pruebas Saber 5°?

1.3 Delimitación del Problema

La presente investigación se enfocó en los resultados de Pruebas Saber 5° de los años 2014, 2015 y 2016 a nivel nacional, los cuales fueron obtenidos de las bases de datos del Instituto Colombiano para la Evaluación de la Educación (ICFES). De estos se buscará los factores asociados al desempeño académico de los estudiantes mediante Minería de Datos.

1.4 Viabilidad de la Investigación

La investigación resultó viable puesto que se tuvo acceso a las bases de datos de los resultados de las Pruebas Saber 5° proporcionadas por el Instituto Colombiano para la Evaluación de la Educación (ICFES) ya que son documentos abiertos al público.

1.5 Limitación de la Investigación

Una limitación fue la fidelidad y veracidad de las bases de datos ya que al ser depuradas por el investigador se pudo perder información relevante para el estudio de los factores asociados.

1.6 Justificación

Tras la revisión de literatura sobre educación y específicamente en rendimiento escolar se ha encontrado que existen estudios como el de Gonzáles, C., Caso, J., Díaz, K. y López, M. (2012), Erazo, O. (2012), Chica, S., Galvis, D. y Ramírez, A. (2010) concluyentes en que el rendimiento académico está asociado no solo a factores internos de las instituciones educativas, por ejemplo, el número de estudiantes por docente, la infraestructura de la institución, los recursos tecnológicos, por mencionar algunos; sino que éste es subjetivo, puesto que, interviene el contexto donde se encuentran los estudiantes, el grado de escolaridad de los padres, su nivel socioeconómico, la distancia desde su lugar de residencia y la escuela, el grado de afectividad de sus familiares, entre otros. Además de estos estudios también se encuentran otros (Fernández, H. 2005), en los que hacen evidenciar la gran preocupación que agentes del estado muestran sobre el rendimiento de los estudiantes en las pruebas por su posicionamiento en relación a otros países. De aquí que se estén tomando medidas para el diseño de políticas educativas que busquen el mejoramiento de los aprendizajes de los estudiantes y por ende mejores resultados en las pruebas externas como las SABER.

En efecto, es relevante este estudio ya que tiene dos implicaciones prácticas; primera, la determinación de los factores socioeconómicos, académicos e institucionales asociados al rendimiento académico de estudiantes de grado quinto en las Pruebas Saber 5; y, la segunda, producir conocimiento que sirva de soporte al MEN, al ICFES y demás entes estatales, como también a la comunidad educativa y a la comunidad en general, para la toma de decisiones tendientes a mejorar la calidad de la educación en el país, y en particular, para ajustar la política pública de la educación básica, media y superior en Colombia. Además, el conocimiento generado en este proyecto se convierte en un aporte importante, de la Universidad Tecnológica de Pereira y

de la Universidad de Nariño, al conocimiento existente a nivel nacional, sobre el rendimiento académico de los estudiantes de educación básica en las Pruebas Saber 5° y a las nuevas formas de investigar este problema, utilizando la Minería de Datos.

En este orden de ideas se establece como objetivo principal descubrir los factores asociados al desempeño en las Pruebas Saber 5° con técnicas descriptivas de Minería de Datos. Además de esto, se tendrá en cuenta los resultados de pruebas en los años 2014 al 2016, periodo en el cual no se han encontrado estudios sobre pruebas saber con relación a la técnica que se empleará, esto permitirá analizar en ellas los cambios que se han hecho desde el ICFES con la alineación[1] de las pruebas teniendo como referencia las competencias genéricas.

1.7 Antecedentes del Tema

El rendimiento académico de los estudiantes es un tema que ha sido objeto de estudio de diversos autores; este es entendido como el proceso de aprendizaje dependiente de diversas variables tanto objetivas como subjetivas, las primeras hacen referencia a los sistemas de evaluación, las calificaciones, las políticas educativas, entre otras. Y las variables subjetivas son las que involucran aspectos cognitivos, familiares, sociales y socioeconómicos. (Erazo, O. 2012).

Dichas variables o factores asociados al rendimiento académico de los estudiantes han sido un tema de preocupación que data desde los años 60's empezando con el estudio de Coleman et al (1966), en el cual se concluyó que el rendimiento escolar en los Estados Unidos estaba influenciado en gran medida por las características socioeconómicas de los estudiantes y por el hecho que, poco o nada tenían que ver en el desempeño académico, las variables asociadas a la institución educativa. Estos resultados generaron controversia, ya que muchos críticos del tema no concebían

[1] La alienación de las pruebas de evaluación de competencias genéricas de SABER PRO representan los eslabones finales de unas series de pruebas que se aplican desde la educación básica.

que las variables asociadas a la institución educativa no tuvieran influencia en el desempeño académico.

A nivel internacional se destaca el estudio realizado por (Gonzáles, C., Caso, J., Díaz, K. y López, M. 2012), titulado *Rendimiento académico y factores asociados. Aportaciones de algunas evaluaciones a gran escala*. En el cual se hace el análisis de factores asociados al rendimiento académico como parte de la aplicación de evaluaciones a gran escala, en las cuales son determinantes algunas variables que van desde el nivel del sistema educativo hasta el nivel del estudiante, como es el caso de las pruebas internacionales PISA, cuyo propósito es evaluar el grado de desarrollo de las competencias básicas que manifiestan los jóvenes de los diferentes países. También existen pruebas internacionales que tienen en cuenta cuatro áreas a saber: el contexto nacional y comunitario, el contexto escolar, el contexto del aula y las características y actitudes de los estudiantes, que se evidencia en las pruebas TIMMS, la cuales pretenden identificar qué se espera que los estudiantes aprendan, cómo se organiza la enseñanza y en qué contexto ocurre. Además de estas pruebas existen pruebas nacionales como las que se aplican en México, que intentan delimitar lo que los estudiantes mexicanos en su conjunto aprenden del currículo nacional a lo largo de la educación básica (INEE) o aquellas en las que además de tener en cuenta el currículo, parten de otras consideraciones como los programas oficiales de estudio (ENLACE).

Las cuatro pruebas en mención toman en consideración variables asociadas al aprendizaje de los estudiantes y cuyo propósito es recopilar información centrada en caracterizar el sistema educativo. Sin embargo, según los autores es conveniente plantear otro tipo de evaluación más completa que tenga en cuenta por un lado la evaluación de estudiantes en forma censal y por otro lado que mida otras variables que son de gran relevancia como el docente, el aula, el propio individuo y su familia y que no se tiene en cuenta en las pruebas antes mencionadas.

En Colombia se han realizado varios estudios que buscan determinar los factores que influyen en el rendimiento académico de los estudiantes. En el estudio efectuado por Gaviria y Barrientos (2001b), los autores analizaron los resultados de las pruebas de estado de 1999, en el cual, encontraron que las características asociadas a la institución educativa afectan de manera importante el rendimiento académico y que lo hacen en mayor medida que las variables socioeconómicas; además, en él, se reconoce que el nivel educativo de los padres tiene un efecto importante en el desempeño académico. Encontraron, además, que existe una brecha importante entre los resultados de instituciones oficiales y privadas. Estos hallazgos ponen en cuestión los resultados de Coleman et al (1966), en lo atinente a la influencia de las variables institucionales en el desempeño académico de los estudiantes.

Otro estudio realizado por Chica, S., Galvis, D., Ramírez, A. (2010). En el cual identificaron los determinantes del rendimiento académico en Colombia utilizando los resultados obtenidos por los estudiantes en las áreas de matemáticas y lenguaje de las pruebas ICFES Saber 11º del semestre B de 2009, Para esto, se utilizó el modelo Logit Ordenado Generalizado. Los resultados obtenidos con este método enseñan la relevancia que tienen las variables socioeconómicas en el desempeño para las dos áreas.

Una investigación, contratada por *el Instituto Colombiano para la Evaluación de la Educación* (ICFES, 2011), y desarrollado por Luis Jaime Piñeros. En ésta, se presentaron los principales hallazgos del estudio de factores asociados a los resultados de los estudiantes en las Pruebas Saber 5° y 9° aplicadas en 2009. Con el fin de contribuir a lograr una mejor comprensión de aquellos aspectos de los contextos personales, familiares y escolares que inciden en los desempeños de los estudiantes en las pruebas, y aportar a la toma de decisiones de políticas orientadas a mejorar la calidad y la equidad de la educación en Colombia.

Por otra parte, Gutiérrez, Y. (2015), analizó la relación que se presenta en la estructura familiar de los estudiantes de grado 3° y 5° de primaria con el rendimiento académico en el área de Matemáticas en las Pruebas Saber del año 2013. Los resultados obtenidos de este estudio permitieron concluir que el factor asociado que tiene mayor incidencia sobre el desempeño académico de los estudiantes es el tamaño de la familia.

Ahora bien, teniendo en cuenta que la Minería de Datos en la educación no es un tema nuevo y que su estudio y aplicación ha sido muy relevante en los últimos años, se puede utilizar sus técnicas para explicar y predecir fenómenos dentro del campo educativo (Timarán, Calderón y Jiménez, 2013a) y (Timarán, Calderón y Jiménez, 2013b). Por ejemplo, utilizando las técnicas de Minería de Datos, se puede predecir, con un porcentaje muy alto de confiabilidad, la probabilidad de deserción de cualquier estudiante (Valero, 2009) y (Valero, Salvador y García, 2010). Así pues, las instituciones educativas pueden aplicar la Minería de Datos para efectuar análisis comprensivos sobre las características de los estudiantes, de los métodos evaluativos, con lo cual se pueden develar procesos exitosos, como también, detectar fraudes o inconsistencias (Orea, Vargas, y Alonso, 2005), que de otro modo podrían permanecer ocultos.

Recientemente, se ha incrementado el interés en utilizar la Minería de Datos en la educación, centrándose en el desarrollo de métodos de descubrimiento que utilicen los datos de plataformas educacionales y en el uso de esos métodos para comprender mejor a los estudiantes y el entorno en el que aprenden. Los métodos empleados en la Minería de Datos en la educación suelen diferir de los métodos más generalistas, explotando explícitamente los múltiples niveles de jerarquía presentes en los datos (Jiménez, A. & Álvarez, H, 2010).

Ahora bien, teniendo en cuenta que los anteriores estudios han sido un gran aporte a la educación, puesto que ayudan a comprender mejor los factores asociados al rendimiento escolar

mediante el uso de técnicas estadísticas, que han sido frecuentemente utilizadas en este tipo de estudios. Se ha optado por el desarrollo de esta investigación que busca implementar técnicas descriptivas de Minería de Datos con el fin de detectar patrones de rendimiento académico a partir de los resultados de la aplicación de las Pruebas Saber 5 en los años 2014 a 2016; ya que esta técnica permitirá obtener resultados más precisos y que puedan ser una herramienta importante a la hora de la toma de decisiones para el mejoramiento de la calidad educativa.

Cabe mencionar que este será uno de los primeros estudios que se desarrollará a nivel nacional utilizando la Minería de Datos con la metodología CRISP - DM.

1.8 Objetivos

1.8.1 Objetivo General

Descubrir factores asociados al desempeño académico en las competencias que evalúan las Pruebas Saber 5° de los estudiantes pertenecientes a instituciones educativas colombianas, a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES en el periodo 2014 a 2016, a través de técnicas descriptivas de Minería de Datos, que permitan generar conocimiento encaminado a soportar las decisiones institucionales y gubernamentales para el mejoramiento de la calidad educativa.

1.8.2 Objetivos Específicos

- ✓ Apropiar el conocimiento sobre las competencias evaluadas en las Pruebas Saber 5°. Identificar y seleccionar de las bases de datos del ICFES los datos socioeconómicos, académicos e institucionales de los estudiantes del país que presentaron las Pruebas Saber 5° en los años 2014 a 2016.

- ✓ Construir un repositorio inicial de datos con los valores de los atributos socioeconómicos, académicos e institucionales de los estudiantes del país que presentaron las Pruebas Saber 5°, a partir de las bases de datos del ICFES.
- ✓ Aplicar técnicas de limpieza y transformación al repositorio inicial de datos con el fin de obtener datos limpios, correctos, consistentes y discretizados de las Pruebas Saber 5°.
- ✓ Aplicar las técnicas descriptivas de Minería de Datos más apropiadas para el descubrimiento de factores asociados al desempeño académico en las Pruebas Saber 5°, utilizando una herramienta de Minería de Datos de software libre.
- ✓ Evaluar e interpretar los patrones obtenidos con el fin de determinar el conocimiento descubierto acerca de los factores socioeconómicos, académicos e institucionales asociados al desempeño académico de los estudiantes que presentaron las Pruebas Saber 5° a nivel nacional.
- ✓ Visibilizar los resultados de la investigación en un informe final para la posterior sustentación.

2. Conceptualización de Pruebas Saber y Minería De Datos

2.1. Elementos conceptuales de las Pruebas Saber

El Instituto Colombiano para la Evaluación de la Educación (ICFES), siendo una entidad adscrita al Ministerio de Educación Nacional (MEN) la cual está encargada de promover y evaluar la educación colombiana, ha venido realizando la evaluación censal para los estudiantes que terminan la educación media desde 1968, con la aplicación de las pruebas de estado para el ingreso a la educación Superior.

En los años 90, se obtuvo información relevante a través de la aplicación de pruebas muestrales[2] a los grados 3°, 5°, 7° y 9° sobre las competencias básicas en las áreas de Lenguaje y Matemáticas llamadas Pruebas Saber.

Las Pruebas Saber son una herramienta que permite evaluar la calidad de la educación en todos los establecimientos educativos del país tanto públicos como privados. El propósito de aplicar estas pruebas es monitorear el rendimiento académico, entendido como el desarrollo de competencias básicas de los estudiantes y hacer seguimiento a la calidad educativa que se brinda en las instituciones educativas. Además, valorar los avances en un determinado lapso de tiempo, establecer el impacto de los programas y finalmente plantear acciones específicas de mejoramiento.

[2] Para llevar a cabo estas evaluaciones periódicas, y garantizar su confiabilidad y validez, fue construida una muestra maestra representativa de la población estudiantil y de las instituciones educativas – oficiales y urbanas, oficiales rurales y no oficiales – a nivel nacional, departamental y de ciudades capitales como Bogotá, Barranquilla, Cali y Medellín. Esta muestra maestra incluyó 36 entidades territoriales, 31 de ellas correspondientes a calendario A y 5 a calendario B. (Fernández, H. 2005)

Estas pruebas fueron tomando más relevancia a partir del año 1994 puesto que su aplicación se hizo de manera censal para que la información obtenida de ellas fuera beneficiosa para todas las instituciones educativas del país.

Las Pruebas Saber evalúan las competencias de Lenguaje, Matemáticas, Ciencias Naturales y Ciudadanas que han desarrollado los estudiantes hasta quinto grado de Básica Primaria y hasta noveno grado de Básica Secundaria. Su diseño está alineado con los Estándares Básicos de Competencias establecidos por el Ministerio de Educación Nacional entendidos como referentes comunes a partir de los cuales es posible establecer qué tanto los estudiantes y el sistema educativo en su conjunto están cumpliendo con unas expectativas de calidad en términos de lo que saben y saben hacer.

Sumado a esto, las pruebas permiten realizar estudios sobre las variables que más influyen en el rendimiento académico y el ICFES con el fin de mejorar la calidad de la educación del país, extendió el proceso de evaluación e inició el estudio de los factores asociados al rendimiento escolar, para este fin se aplicó un cuestionario de prueba en el año 2012 para recoger información sobre el contexto de los estudiantes, sus familias y la institución educativa. Para este estudio el ICFES implementó el modelo Contexto, Insumos, Procesos y Productos (CIPP) el cual permite seleccionar variables de un contexto particular para describir de forma clara y sencilla como éstas influyen en el rendimiento académico. Ver figura 1.

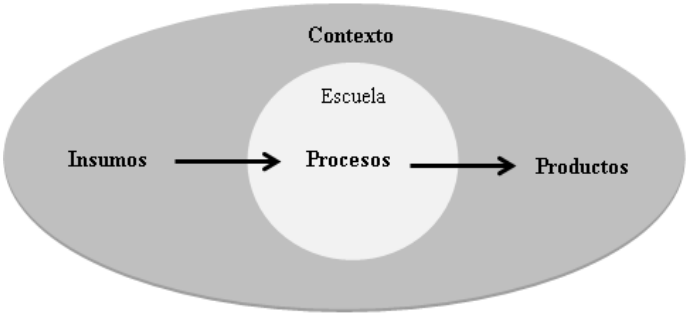

Figura 1 Modelo CIPP. Fuente Marco de factores asociados Saber 3°, 5° y 9° de 2016

En el contexto se recogen todas las variables externas a la institución educativa que y se caracterizan por los aspectos: social, económico, político, cultural y familiar. Dentro de estos se incluyen la ubicación geográfica de la institución educativa, así como la modalidad (escolarizada, formal convencional), dependencia administrativa y tamaño del establecimiento. Otros aspectos relacionados con las características personales de los estudiantes y de sus hogares como el nivel socioeconómico, el género, la motivación y el autoconcepto académico también hacen parte de esta categoría. (ICFES, 2016)

Los Insumos o recursos con los que cuenta la institución educativa incluyendo el historial académico de los estudiantes como un factor determinante de los procesos de enseñanza y aprendizaje. Dentro de los indicadores de infraestructura escolar están los servicios básicos, el acceso a computadoras y la conexión a internet, entre otras, el tiempo efectivo de aprendizaje y las estrategias de clasificación de estudiantes al interior de las sedes; las variables asociadas a los antecedentes escolares como la asistencia a educación preescolar y la repetición de grado también se incluyen en esta categoría.

En los Procesos se tiene en cuenta las actividades y estrategias que se implementan en las instituciones educativas para que los estudiantes apropien los conocimientos y logren desarrollar las competencias que se establecen desde el MEN. Además, se encuentran aquellas variables que permiten caracterizar el clima al interior del aula y de la institución educativa, la gestión de los directores y la satisfacción de los docentes y de todo el personal de la institución.

Finalmente, los Productos son los resultados de las actividades empleadas en las instituciones y las políticas públicas, uno de esos resultados es el desarrollo cognitivo, social, emocional y ciudadano que conllevan a formar una persona integral.

El modelo CIPP cumple cuatro criterios generales que permiten hacer un juicio valorativo; es completo puesto que incluye todas las categorías de los factores asociados al aprendizaje y la influencia de estas en el proceso educativo. Es claro, porque presenta resultados que se pueden entender a partir cualquier ámbito desde especialistas en investigación educativa, así como tomadores de decisiones, docentes y rectores. Presenta direccionalidad de las relaciones que se espera encontrar a partir del planteamiento de hipótesis sobre los factores asociados al aprendizaje. Por último, el modelo es flexible y permite incluir variables que son relevantes para el aprendizaje y desarrollo integral de los estudiantes, esto debido a que se pueden presentar cambios o transformaciones socioculturales.

2.1.1 Factores Asociados

Como se mencionó anteriormente el marco de factores asociados sigue el modelo CIPP, como se muestra en la figura 2, y dentro de cada una de las categorías incluidas en él se definen unas variables que han sido conceptualizadas y analizadas a nivel mundial y pueden usarse de forma dinámica para responder a políticas públicas o adaptarse a cualquier contexto a nivel nacional, regional y local. Para este estudio se tomó como referencia el Marco de factores asociados Saber 3°, 5° y 9° publicado por el ICFES en el año 2016 y se definen a continuación (ICFES, 2016):

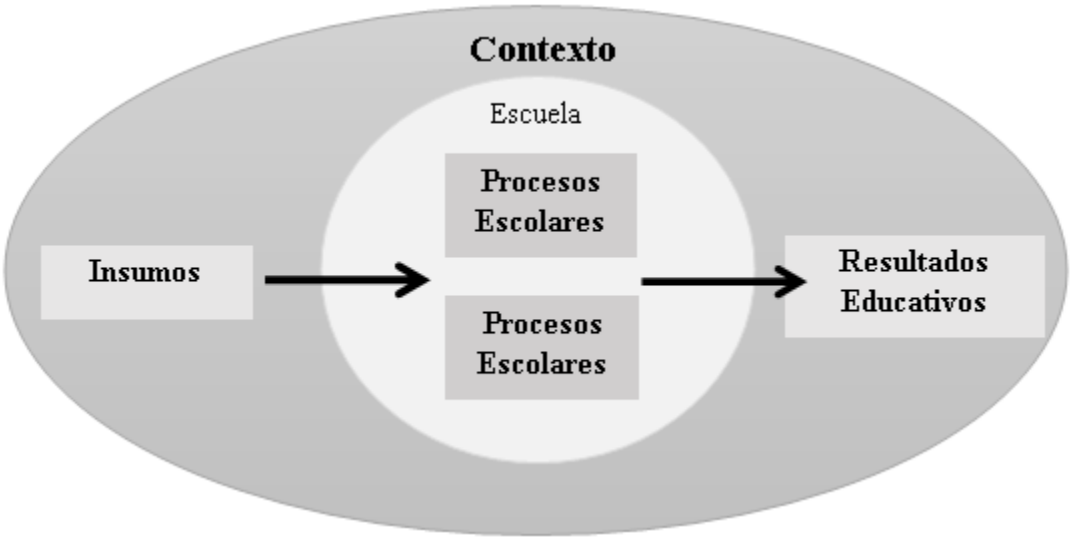

Figura 2 Marco de Factores Asociados. Fuente Marco de factores asociados Saber 3°, 5° y 9° de 2016

2.1.1.1 Contexto

Los factores asociados incorporan las variables más cercanas al proceso educativo porque son estas las que caracterizan a los estudiantes e instituciones educativas, algunas de ellas son:

2.1.1.1.1 Características de los Estudiantes

Según estudios de Coleman (1966) y la UNESCO (2010), los aspectos sociodemográficos, económicos y culturales de los estudiantes y sus familias son variables importantes a la hora de explicar los resultados escolares. Estos factores a pesar de influir en el aprendizaje no se pueden modificar puesto que no dependen las instituciones educativas. Las variables que se consideran de mayor relevancia dentro del marco de factores asociados son:

Género: Según la Unesco, 2010 y 2016a; Goldin et. al., 2006 y Niederle, et al., 2010 se evidencia una brecha de desigualdad en el rendimiento académico de los niños y niñas puesto que se ven reflejados los estereotipos de género que se marcan desde la familia y la sociedad, por ejemplo, la gran desventaja que se muestra en las niñas frente al aprendizaje de las matemáticas y a los niños en lectura.

Edad: Esta variable se relaciona con el grado que cursa el estudiante; permite identificar aspectos como la repetición de año, la extra edad, la desescolarización parcial o deserción escolar.

Según la ley general de educación un estudiante se encuentra en extra edad si supera en dos años la edad promedio del grado cursado; por tanto la escolarización es obligatoria entre los 5 y 15 años de edad, desde primero a noveno grado (ICFES, 2016).

Trabajo Infantil: Es una variable que afecta de manera negativa al rendimiento escolar ya que las condiciones físicas de los niños y niñas disminuyen, pues no están preparados para asumir tareas laborales pesadas.

Se debe diferenciar entre el trabajo infantil y la colaboración en tareas hogareñas puesto que el primero afecta negativamente el rendimiento escolar y el segundo muestra una relación positiva (Unesco 2015a; Unesco 2010).

Nivel Socioeconómico Familiar: Según Célis, Jiménez & Jaramillo (2012) e Icfes (2011), la variable socioeconómica es explicativa frente al desempeño de los estudiantes en las Pruebas Saber, puesto que es una limitante para su desarrollo. Las condiciones del entorno donde viven, el nivel educativo de los padres, la situación laboral, la disposición de artefactos tecnológicos, entre otras, son algunos aspectos de interés ya que son un aliciente importante para el rendimiento escolar.

Violencia en el entorno del hogar: Un entorno violento tiene influencias negativas en cuanto al rendimiento escolar y la formación ciudadana (Chaux, 2009). Para poder determinar el efecto que tiene la violencia en el entorno familiar y en el desempeño académico se debe hacer un estudio de la frecuencia con que ocurren dichas situaciones y en qué tipo de violencia se está incurriendo.

Involucramiento Parental: El acompañamiento y participación de los padres en el proceso de formación académica beneficia el desarrollo cognitivo y no cognitivo de los hijos. Crea actitudes de motivación frente al aprendizaje, reflejadas en los resultados escolares, se fortalece el desarrollo motivacional y ayuda a crear hábitos de estudio en familia (Pomerantz, Moorman & Litwack, 2007).

Pertenencia a un Pueblo Originario o Etnia: El desafío que hay en las instituciones educativas es el de incluir y atender a estudiantes que pertenecen a alguna etnia en particular; ya que estos presentan condiciones socioeconómicas bajas y de difícil acceso a la educación. Por esto, los establecimientos educativos deben apropiar sus currículos para la flexibilización del aprendizaje y las necesidades que son requeridas por los estudiantes (Unesco, 2015b).

Necesidades Educativas Especiales: La política de una "Educación para todos" planteada por la UNESCO deja ver la importancia de tener en cuenta esta variable NEE dentro de los currículos de los establecimientos educativos del país, ya que incide significativamente en el desempeño académico de los estudiantes. Se deben distinguir entre necesidades educativas cognitivas y físicas, dado que en las primeras se debe buscar fortalecer y alcanzar los logros mínimos planteados para cada grado, mientras que las necesidades físicas se asocian más a la adecuación de espacios físicos que les permita la accesibilidad al establecimiento (Unesco, 2005).

Auto-Concepto Académico: Es entendido como las creencias que tienen los estudiantes sobre sus capacidades para hacer bien sus labores académicas y obtener buenos resultados. La confianza que tenga cada estudiante sobre su desempeño afecta de manera positiva o negativa sus resultados; esto se evidencia en el cumplimiento de tareas, la participación en clases y la asistencia a las mismas.

En la formación del auto-concepto académico se tiene en cuenta aspectos como la comparación personal, la comparación social, la percepción de los padres, profesores y compañeros (Kurtz-Dostes y Scheneider, 1994)

Motivación: Esta variable muestra una estrecha relación con el desempeño académico de los estudiantes es así como Ryan and Deci, (2000) proponen dos tipos de motivación: la intrínseca cuando se refiere a la motivación causada por gusto e interés propio y la extrínseca cuando es causada por algún incentivo en particular. En el marco conceptual de los cuestionarios PISA, la motivación es reconocida como parte de los comportamientos o actitudes positivas frente al aprendizaje (OECD, 2003; OECD, 2010b; OECD, 2016a) y cuando se logra controlar la variable socioeconómica la motivación intrínseca tiene resultados positivos frente al aprendizaje.

Además, la motivación en el ámbito educativo se debe entender como la manera en que las relaciones entre los estudiantes y sus profesores entran en juego para facilitarla o limitarla con el fin de lograr un aprendizaje significativo, razón por la cual el profesor debe favorecer la denominada motivación intrínseca, que lleva a la curiosidad y el descubrimiento de lo nuevo, a diferencia de la motivación extrínseca en la que el estudiante es movido por otros para realizar diferentes actividades (Ospina J, 2006)

Estrategias de Aprendizaje: Las estrategias de aprendizaje son las que facilitan la adquisición del conocimiento y la adquisición de nuevas habilidades.

Existen diferentes estrategias de aprendizaje pero se intensifican los estudios en las tradicionales y las cooperativas. Caracterizándose las primeras por el uso de la memorización como recurso de aprendizaje y las segundas por la creación colectiva del conocimiento (Weinstein, Husman, & Dierking, 2000).

2.1.1.1.2 Características de las Escuelas

En este apartado se analizan aquellas variables que hacen referencia a las instituciones educativas. Dichas características están referenciadas en los cuestionarios que se aplican a los estudiantes al momento de presentar las Pruebas Saber. Entre las variables más relevantes en esta sección están:

Zona en la que se Ubica la Institución Educativa (Urbana o Rural): Sobre esta variable se han realizado estudios que evidencian que las instituciones que se encuentran más aisladas poseen menos posibilidades de contar con docentes idóneos, materiales y recursos tecnológicos, apoyo técnico y se da una mayor deserción estudiantil, principalmente por la falta de recursos económicos de los padres que en muchas ocasiones emplean a sus hijos para realizar las labores del campo. Además de esto, en diversos establecimientos que se encuentran en las zonas rurales existen modos de organización diferentes pues es común encontrar escuelas; multigrado o unidocente (ICFES, 2016).

En este sentido la ubicación de la institución educativa es un factor que cobra importancia para explicar el desempeño académico de los niños y niñas del país y se puede medir diferenciando entre escuelas rurales y urbanas, el municipio, el departamento, la zona y la entidad territorial en la que se encuentra.

Dependencia Administrativa de la Escuela: La influencia ejercida por el tipo de dependencia es una de las variables que cobra mayor importancia dentro de las características de las escuelas, ya que algunos autores como Chubb (2001) consideran que los entes privados exigen por parte de los directivos y docentes el mejoramiento continuo en los aprendizajes de los estudiantes por la presión ejercida por los padres. Sin embargo en la práctica se ha encontrado que

dicha afirmación en pocos casos se cumple, indicando que la educación privada por sí sola no asegura mejores aprendizajes y oportunidades para los estudiantes (Ball, 2012).

Es importante dentro de este aspecto distinguir dos dependencias administrativas: públicas y privadas, las primeras son controladas por el estado, las segundas están regidas por algún ente privado, además se pueden diferenciar por la orientación religiosa, la procedencia de los aportes o los llamados colegios en concesión donde se entrega la administración de un colegio que cuenta con infraestructura pública a uno privado.

Tamaño de la Institución Educativa y el Aula de Clases: Frente a este factor existen contraposiciones de diversos autores frente al tamaño de la institución y el número de estudiantes en el aula; Crenshaw (2003) y Lamdin (1995) afirman que la cantidad de estudiantes atendidos en la institución educativa influye en el tipo de enseñanza y en el proceso de aprendizaje de los estudiantes, sin embargo Ajani & Akinyele (2014) y Stevenson (1996) se enfocan más en el tamaño relativo de la institución afirmando que el número de docentes por estudiante explica de mejor manera los efectos en el apoyo del proceso de enseñanza.

Otros factores que se tienen en cuenta para explicar este factor hacen referencia a la capacidad de la planta docente, por la cantidad de cursos y niveles, y por el tipo de educación: técnica profesional, artística, científica, entre otros.

Nivel Socioeconómico Promedio por Aula y Establecimiento: Este factor se puede determinar a partir del estrato de la institución y toma importancia al identificar el efecto de los pares en cada estudiante, el cual es beneficioso en el desarrollo escolar pues se da la heterogeneidad social y académica de los estudiantes, a nivel de aula y escuela, además impacta el desarrollo del currículum, la calidad de las interacciones y la capacidad del profesor para desarrollar procesos de enseñanza efectivos (Duflo, E. et al, 2011). Algunas características asociadas al nivel

socioeconómico tienen que ver con el ingreso, posesión de bienes, nivel de estudios de la madre y/o padre, ocupación de la madre y/o padre.

Violencia en el Entorno de la Institución Educativa: Según Bowen & Bowen (1999) y la Unesco (2015b) esta característica evidencia las situaciones de conflicto y por ende de procesos de marginación al interior de las instituciones educativas. Las repercusiones al respecto recaen principalmente en los estudiantes, frustrando su adecuado desarrollo académico y posteriormente la bajos desempeños en Pruebas Saber.

2.1.1.2 Insumos

Los insumos o recursos educativos con los que cuenta la institución educativa son una parte importante en los resultados académicos; ya que a través de estos los estudiantes adquieren un aprendizaje significativo que fortalece sus competencias escolares. A continuación se presentan aspectos que se relacionan con esta categoría:

Antecedentes Escolares: Es importante reconocer que los estudiantes traen aprendizajes y conocimientos previos que influyen de manera directa en el rendimiento académico. Factores asociados como la participación en preescolar y la repetición de grado son antecedentes que traen consigo los estudiantes y se consideran favorables para la adquisición de nuevas competencias en la formación académica (ICFES, 2016).

Asistencia a Educación Preescolar: Según la OECD (2012) la educación preescolar tiene efectos positivos sobre los resultados escolares y en otros ámbitos como la reducción de la pobreza y la movilidad social; por tal razón la asistencia al preescolar es favorable para el desarrollo de la sociedad en general. En este sentido, los gobiernos deben enfocarse en desarrollar programas que fortalezcan y apoyen este grado buscando equidad, calidad y cumplimiento con los estándares mínimos que les permitan a los niños y niñas gozar de todos los beneficios que se puedan ofrecer

en pro de la calidad en el aprendizaje. Además de esto, la vinculación a programas de educación preescolar ofrece características diferenciadoras en las trayectorias escolares de los niños y niñas; quienes asisten a educación preescolar formal llegan mejor preparados para ajustarse a las exigencias escolares de la educación básica y media ofrecida en los distintos establecimientos del país.

Repetición de Grado: Este tema ha sido estudiado por diferentes autores como Holmes (1989) y Roderick (1994) quienes indican que la repetición de año tiene efectos negativos a nivel académico y emocional a través del tiempo. Cuanto menor sea el grado de repetición mayores serán los efectos o consecuencias negativas para el aprendizaje. En Latinoamérica, la repitencia de grado en la institución educativa ha sido identificada como la variable que más influencia en los logros de aprendizaje, luego del nivel socioeconómico de los estudiantes, lo cual refleja un impacto negativo como medida efectiva para mejorar el desempeño académico (Unesco, 2015b). En la actualidad Colombia es uno de los países de Latinoamérica con las tasas de repitencia más bajas en educación primaria; sin embargo, al considerar las consecuencias de esto, resulta conveniente revisar con mayor detalle las políticas del sistema educativo colombiano (Unesco, 2015a).

Uso de Tecnologías: Hoy en día la sociedad esta movida por las Tecnologías de la Información y la Comunicación (TIC´s) pues brindan múltiples posibilidades de acceder a información de manera más rápida y efectiva, razón por la cual incluir dentro de la práctica docente el uso de las TIC´s, reducidas al uso de computadoras, dispositivos electrónicos y el acceso a internet, es de vital importancia dentro de los procesos de enseñanza y aprendizaje. Uno de los aportes de las TIC´s es el aprendizaje colaborativo que puede darse; a nivel interinstitucional entre docentes y estudiantes y a nivel intrainstitucional facilitando la colaboración y el aprendizaje

colectivo entre familias y centros educativos alrededor del mundo, y permitir el acceso a procesos formativos virtuales para todos los ciudadanos Gómez y Macedo (2010).

Según los resultados obtenidos en las pruebas PISA 2012 se evidencia pequeños efectos en los resultados, situación que se explica por la poca formación de docentes en cuanto a las capacidades de entendimiento y pensamiento frente a estrategias con el uso de las tecnologías y de estudiantes y sus inexistentes prácticas de pedagogía encaminadas hacia uso intensivo y adecuado de las TIC (OECD, 2015).

Infraestructura y Equipamiento Escolar: La relación que existe entre la infraestructura y el rendimiento académico presenta una relación positiva ya que; disponer de una biblioteca, de computadores o tecnologías de información y comunicación, de espacios recreativos, de salones de clase adecuados para todos los estudiantes y de servicios sanitarios adecuados, le da la posibilidad a los estudiantes de contar con ambientes propicios que generan mayor interés y motivación por el aprendizaje (Unesco, 2015b).

Calificación Docente y Conocimiento Profesional: Un buen docente se caracteriza por tres aspectos que lo hacen idóneo en su desempeño laboral: los criterios de selección inicial, la calidad de instrucción y la capacitación continua.

Además, es importante que el docente conozca las características de sus estudiantes y busque estrategias pedagógicas que le certifiquen el aprendizaje de sus alumnos independientemente de sus características de origen (Barber & Mourshed, 2008).

Materiales Educativos: Hace referencia a los recursos o herramientas que permiten la profundización de los conceptos y la apropiación del conocimiento tanto de estudiantes como docentes.

Algunos elementos que hacen parte de los materiales educativos son los artefactos tecnológicos, los libros de texto, la disponibilidad de internet, los útiles escolares, entre otros. El uso y aprovechamiento que se les dé a estos recursos en el aula de clases durante el desarrollo de las sesiones pedagógicas define en gran medida los resultados esperados frente al logro educativo (ICFES, 2016).

Programa de Útiles Escolares Gratuitos: Si bien es cierto este tipo de programas son de vital importancia para apoyar los procesos de aprendizaje de los estudiantes, ya que les permite tener acceso y disponibilidad de herramientas educativas para la adquisición de conocimientos, también se sabe que estos programas tienen criterios de focalización (colegios o estudiantes más desventajados socioeconómicamente) sin embargo los materiales que entregan a los establecimientos educativos muchas veces no son necesariamente los que se requieren para orientar el aprendizaje, pues como se mencionó anteriormente no se cuenta con la instrucción adecuada para que el uso adecuado de dichas herramientas (Reimers, DeShano daSilva, & Treviño, 2006).

Tiempo Efectivo Dedicado al Aprendizaje: Este componente hace referencia al tiempo que se dedica exclusivamente al aprendizaje haciendo diferencia entre las horas de estudio de los estudiantes y el tipo de jornada al que asisten. Otros factores que permiten identificar el tiempo efectivo dedicado al aprendizaje según Murillo (2007) son la inasistencia escolar y la puntualidad por parte de estudiantes y docentes, la frecuencia con la que ocurren interrupciones durante la clase, debido a un ambiente de indisciplina nocivo para el aprendizaje o a factores de distracción externos al control del docente como ruido o condiciones climáticas inapropiadas, y el manejo y organización del tiempo por parte del docente.

2.1.1.3 Procesos

Los procesos que ocurren en la institución educativa y en los salones de clase representan el núcleo de la labor educativa, puesto que a través de las interacciones cotidianas entre docentes y estudiantes se promueve el aprendizaje (ICFES, 2016).

Los constructos que hacen parte de los procesos educativos se definen a continuación:

Liderazgo Educativo: Esta es una de las variables que más influye en el aprendizaje de los estudiantes dado el papel que juega el desarrollo de programas, proyectos y prácticas educativas. Los roles de liderazgo educativo orientan principalmente al logro del aprendizaje y se focalizan en los aspectos de la gestión instruccional. El liderazgo instruccional es un factor importante para orientar las metas y focalizar los agentes internos hacia los resultados académicos en los procesos de enseñanza. Algunos de los procesos escolares transformadores que deben ser tomados en cuenta por las personas a cargo del liderazgo educativo son: empoderamiento de los profesores, fortalecimiento de la visión institucional y de objetivos compartidos, y capacidades de rendición de cuentas (Leithwood, 1994).

Desarrollo y Colaboración Profesional Para Mejorar la Enseñanza: El trabajo cooperativo y colaborativo entre maestros permite mejorar el aprendizaje y el desempeño escolar; el intercambio de experiencias sobre las prácticas de aula, conlleva a mejorar las propias y contribuye a desarrollar habilidades en los estudiantes.

Además de esto, el trabajo colaborativo permite fortalecer la planificación, el desarrollo, la evaluación y la retroalimentación para mejorar las prácticas educativas (Elmore, 2010), (Unesco, 2010), (OECD, 2016a).

Clima Escolar: Hace alusión al conjunto de normas y acuerdos que se establecen en la institución educativa para propiciar un ambiente acorde a las necesidades de los estudiantes y

garantizar un aprendizaje efectivo. Un clima laboral y escolar positivo se asocia con los logros académicos (OECD, 2016a). Algunos constructos que se relacionan con esta variable son:

Presencia Efectiva de Normas y Acuerdos: Un ambiente de aprendizaje armonioso, basado en el respeto con normas claras y el cumplimiento de las mismas, potencia el aprendizaje estudiantil; contrario a esto, cuando se presentan ambientes hostiles, de autoritarismo e irrespeto por parte del docente o los estudiantes, se genera indisciplina y pérdida de tiempo valioso que podría dedicarse a la adquisición de conocimientos (OECD, 2016a).

Relaciones Interpersonales: Entre los aspectos a tener en cuenta en las relaciones interpersonales están la comunicación efectiva y asertiva entre los actores educativos así como las relaciones basadas en la confianza. Es impotente hacer énfasis en la actitud y disposición que tienen los docentes para resolver las dudas de sus estudiantes, la actitud de los estudiantes frente a los procesos de enseñanza y el tipo de comunicación que se de en clase (Cohen, McCabe, Michelli & Pickeral, 2009).

Percepción Sobre el Aula y la Institución Educativa: En la literatura se argumenta que existen creencias frente a las experiencias vividas por los alumnos en la escuela que afecta significativamente su desempeño académico y su comportamiento (Lester, Garofalo & Kroll, 1989; Wang & Holcombe, 2010). En este sentido la práctica docente cobra la mayor importancia y debe estar basada en aspectos como; claridad en los objetivos planteados para el desarrollo de las clases de manera que los estudiantes comprendan el sentido de las actividades planteadas, la escucha de inquietudes y el apoyo frente adversas situaciones que se les puedan presentar, preguntar sobre cómo se sienten

en la institución y propiciar relaciones que permitan favorecer la permanencia y el éxito escolar.

Prácticas de Enseñanza Docente: Este aspecto hace énfasis en la calidad instruccional del profesor en donde se evidencia la interacción entre este y sus estudiantes, generando una relación positiva si se cuenta con un conjunto de habilidades por parte del docente, entre las cuales están en primer lugar el desarrollo de conceptos, la estimulación y el desarrollo de habilidades de pensamiento superior, la integración de conceptos y de conocimientos previos, la producción propia y las relaciones y aplicaciones al mundo real. En segundo lugar, la calidad de la retroalimentación dada por parte del docente a los estudiantes, donde el profesor debe desarrollar bases claras para que el estudiante logre avanzar en su aprendizaje, que haga seguimiento constante, que promueva la metacognición y que reafirme positivamente a los estudiantes. Y en tercer lugar, la modelación lingüística, que se refiere a la capacidad del profesor para extender y desarrollar constantemente el lenguaje en los estudiantes, a través de conversaciones frecuentes, preguntas abiertas y el uso de lenguaje avanzado, entre otros (Pianta, Hamre, & Allen, 2012).

2.1.1.4 Resultados Educativos

Actualmente la expectativa de los sistemas escolares, frente al objetivo de promover el desarrollo integral de los estudiantes, se ha convertido en la base fundamental para propender por un buen desempeño académico; partiendo del aprendizaje en las disciplinas académicas, el desarrollo de competencias ciudadanas y el bienestar subjetivo de los estudiantes. Por esta razón es importante describir cada una de ellas como se hace a continuación.

Aprendizaje en Disciplinas Académicas: Según lo establecido en los estándares básicos de competencias es de vital importancia tener un criterio claro y público que permita juzgar si un estudiante, una institución o el sistema educativo en su conjunto cumplen con unas expectativas

comunes de calidad, a esto se define estándar y partiendo de esta definición cobra importancia la medición de los aprendizajes en las distintas disciplinas académicas de las cuales no solo hacen parte las áreas básicas como matemáticas y lenguaje, sino también aquellas que forman parte del desarrollo integral del estudiante y su formación como ciudadanos (Ferrer, 2006).

Ciudadanía: La formación ciudadana hace referencia, de forma general, a aprender a vivir con otros y a participar activamente de una sociedad democrática. En este sentido, este factor asociado se debe tener en cuenta como primordial en la formación académica y ciudadana de los estudiantes, siendo las instituciones educativas el lugar propicio para la socialización y el desarrollo de habilidades y actitudes en relación a la vida en democracia, el respeto por los otros, el cumplimiento de las normas acordadas, participar y expresar su opinión, entre otro aspectos (Espínola, 2005), (Westheimer & Kahne, 2004).

Bienestar Subjetivo: A pesar de ser un factor muy importante en la mediciones de resultados académicos muchas veces quedan relegadas a un segundo plano, debido a que las variable de tipo psicológico asociadas a aspectos socioemocionales no presentan características generales que permitan una medición; el bienestar subjetivo se puede entender como la percepción de bienestar individual, que hace parte de la inteligencia emocional definida según Coleman D. (1995) como la capacidad que tiene una persona para reconocer los sentimientos propios y ajenos y por lo tanto es inteligente o hábil para el manejo de los sentimientos. En este sentido se debe tener en cuenta este factor dentro del proceso de aprendizaje de los estudiantes.

2.1.2 Estructura y Alineación de las Pruebas Saber

A partir del segundo semestre del año 2014 las pruebas fueron alineadas para establecer comparativos con los diferentes exámenes que se aplican como son las Pruebas Saber 3°, 5°, 9°, Saber 11 y Saber Pro. Esto con el fin de monitorear el progreso que han tenido los estudiantes

después de aplicar la prueba en un determinado grado, por ejemplo, un estudiante de primaria que presentó la prueba Saber 3° después de dos años aplica la prueba Saber 5° con los resultados obtenidos de las dos pruebas se puede establecer un comparativo y hacer un análisis del proceso de formación que alcanzó en la Básica Primaria (ICFES, 2013).

Las Pruebas Saber miden las competencias, es decir, pretenden indagar en los niños cómo piensan y cómo utilizan el saber adquirido en los diferentes contextos de aprendizaje. Se entiende por competencia al conjunto de habilidades, destrezas y conocimientos que desarrolla una persona para comprender, transformar y participar de manera activa y crítica en el mundo que lo rodea.

Actualmente, el Ministerio de Educación Nacional (MEN) concibe el objetivo de la educación como el desarrollo de determinadas competencias y, en consecuencia, a estas como el objeto de la evaluación. Dentro de las diferentes competencias que pueden desarrollarse a lo largo del proceso educativo hay una categoría que merece especial atención: la de las competencias genéricas, entendidas como aquellas que resultan indispensables para el desempeño social, laboral y cívico de todo ciudadano, independientemente de su oficio o profesión. Contrastan con las competencias (no genéricas) propias de oficios o actividades laborales particulares, que resultan de un entrenamiento especializado (ICFES, 2013).

De lo anterior, las competencias genéricas son la base más importante en la formación que reciben los niños y niñas de la educación básica; puesto que no solamente son indispensables para subsistir en la vida cotidiana, sino que también le permite al individuo tener una visión exitosa en cuanto a la participación en el mercado laboral, en la relaciones interpersonales en su contexto, con su familia, en procesos democráticos que aporten un bien para la construcción de la paz.

Desde este punto de vista, el MEN enfocó las competencias genéricas como las competencias que se desarrollan en el área de Lenguaje, Matemáticas y Ciudadanía. De tal manera

que todo ciudadano debe estar en la capacidad de leer de manera comprensiva artículos, noticias, redactar cartas, correos electrónicos, de sacar cuentas al momento de hacer sus compras, de realizar un presupuesto y llevar su economía familiar, además, de ser un ciudadano que conoce sus derechos y deberes y que está en capacidad de elegir un representante con la convicción de conocer y comprender sus propuestas.

Según Rychen y Salganik (2006), una competencia es más que conocimientos y destrezas. Involucra la habilidad de enfrentar demandas complejas, apoyándose y movilizando recursos psicosociales (incluyendo destrezas y actitudes) en un contexto en particular. Por ejemplo, la habilidad para comunicarse de manera asertiva a través del buen uso del lenguaje, de actitudes positivas con otras personas y buenas prácticas comunicativas en su entorno, son aspectos que evidencian la movilización de competencias.

Por otra parte, las competencias genéricas se caracterizan como longitudinales, es decir, que deben desarrollarse a lo largo de todo el proceso educativo sin importar el ciclo de formación y deben de ser transversales, de tal manera que todas las áreas del conocimiento contribuyan a la formación de estas competencias.

En este documento, se entienden las competencias a la manera de Rodríguez (2007), es decir, como un conjunto identificable y evaluable de conocimientos, actitudes, valores y habilidades relacionadas entre sí, que permiten desempeños satisfactorios en situaciones reales y en contextos específicos. En este sentido, una persona demuestra que es competente en la acción y no con la repetición de un saber determinado; en otras palabras, se evidencia la competencia cuando el conjunto de saberes se proyecta en acciones concretas que demandan su ejecución consciente; por lo tanto, las competencias se manifiestan en los desempeños que tiene el estudiante en situaciones específicas. Las competencias reconocen diversos grados de desempeño y de logros,

expresados mediante indicadores, que permiten identificar los diferentes momentos o niveles de logro que constituyen una competencia determinada. Las competencias son dinámicas y se transforman de acuerdo con las experiencias del individuo y con los procesos de aprendizaje (Torrado, M, 2000, p.121).

El ICFES para el año 2009 diseñó las Pruebas Saber 3°, 5° y 9° de tal forma que garanticen evaluaciones censales para un periodo de doce años con el fin de visualizar en los resultados la evolución que ha tenido la educación en Colombia. Estas pruebas evalúan las competencias en Matemáticas, Lenguaje y Ciencias Naturales; cabe resaltar que éstas no permiten evaluar la totalidad de las competencias que deben adquirir los estudiantes durante su instancia en la escuela pero si son un aliciente para que ellos continúen su formación profesional, laboral y social a lo largo de toda su vida (ICFES, 2011).

El diseño de las pruebas fue realizado por docentes expertos en el área y se tuvo en cuenta los Estándares Básicos de Competencias tomados como un referente común en cuanto a lo que deben saber y saber hacer los estudiantes durante su proceso educativo independientemente de su lugar de origen, modalidad educativa, condiciones culturales, sociales, económicas, religiosas, entre otras.

En la tabla 1 se evidencia la conformación de las Pruebas Saber 2009 teniendo en cuenta las competencias en Lenguaje, Matemáticas y Ciencias Naturales.

Tabla 1 Procesos de las competencias genéricas

Lenguaje	Matemáticas	Ciencias Naturales
Lectura **Escritura**	Razonamiento y argumentación Comunicación, representación y modelación Planteamiento y resolución de problemas	Uso comprensivo del conocimiento científico Explicación de fenómenos

		Indagación

Fuente.: Guía de lineamientos generales de Pruebas Saber 2009

Además de esto, en cada una de ellas se evalúan los componentes que son los ejes verticales con los Estándares Básicos de Competencias y que permiten evidenciar las fortalezas y debilidades que tienen los estudiantes como se muestra en la tabla 2.

Tabla 2 Componentes de las competencias genéricas

Lenguaje	Matemáticas	Ciencias Naturales
Semántica	Numérico – Variacional	Entorno vivo
Sintaxis	Geométrico – Métrico	Entorno físico
Pragmática	Aleatorio	Ciencia, tecnología y sociedad (CTS)

Fuente. Guía de lineamientos generales de Pruebas Saber 2009

La Prueba contiene preguntas de selección múltiple con única respuesta, estas presentan un enunciado con cuatro opciones de respuesta A, B, C y D y sólo una de ellas es la correcta, el número de preguntas que contesta cada estudiante y en particular los de grado quinto son: 36 para Lenguaje, 48 en Matemáticas y 48 para ciencias Naturales.

Ahora bien, teniendo en cuenta el Plan Nacional Decenal de Educación 2006-2016, el ICFES ha avanzado en la alineación del Sistema Nacional de Evaluación Externa Estandarizada (SNEE), a través de la reestructuración de los exámenes: en 2009 con un nuevo diseño de Saber 3°, 5° y 9°; en 2010 con el rediseño de Saber Pro; en 2014 con los cambios en Saber 11°. La alineación posibilita la comparación de los resultados en distintos niveles educativos, ya que los

diferentes exámenes evalúan unas mismas competencias en algunas de las áreas que los conforman, a saber, las competencias genéricas (ICFES, 2014).

En el caso particular del examen Saber 5°, que se aplica desde el segundo semestre del 2014, la alineación consistió en la introducción de una prueba de competencias ciudadanas la cual presenta situaciones de análisis que se relacionan con su entorno más cercano, es decir, el aula, el colegio, la familia y el barrio, con menor complejidad y en un lenguaje más sencillo que las de grados 9° (ICFES, 2014).

2.2 Elementos Conceptuales de la Minería de Datos

La Minería de Datos se centra en el descubrimiento de patrones obtenidos al analizar grandes volúmenes de información almacenados en bases de datos. La información contenida en estas se obtiene; por la ejecución de diversas tareas realizadas por las personas en el ámbito industrial o educacional y se recopilan a través de sistemas informáticos o computacionales. Los registros que se manejan en las bases de datos han variado con el tiempo debido a que en otras épocas los sistemas informáticos no presentaban la capacidad que actualmente disponen, un ejemplo claro de ello es la concepción en los años ochenta sobre el número de registros; considerando alrededor de mil registros como una gran cantidad de información. Sin embargo con el paso de los años los sistemas informáticos han evolucionado enormemente hasta el punto de almacenar la información de millones y millones de aspectos desde todos los ámbitos administrativos.

Actualmente, los datos que se generan son innumerables pues la entrada de las computadoras personales y el uso del internet brinda la posibilidad de múltiples conexiones. A parte de las computadoras, existen otros artefactos tecnológicos como son: celulares, televisores digitales, tabletas, entre otros, que han llevado al crecimiento exponencial de los datos.

A manera de ejemplo se puede mencionar el incrementado interés en utilizar la minería de datos en el estudio educacional, centrándose en el desarrollo de métodos de descubrimiento que utilicen los datos de plataformas educacionales y en el uso de esos métodos para comprender mejor a los estudiantes y el entorno en el que aprenden (Jiménez y Álvarez, 2010).

Sin embargo, a partir de la información obtenida en las plataformas educacionales, se queda una gran cantidad de información oculta de gran importancia estratégica, y que no es posible analizar con técnicas meramente estadísticas, sino que se hace necesario el uso de la minería de datos que parte de técnicas de inteligencia artificial para encontrar patrones y relaciones dentro de los datos, llegando al descubrimiento del conocimiento (Knowledge Discovery in Databases, KDD por sus siglas en inglés) (Agrawal y Srikant, 1994) (Chen et al., 1996) (Piatetsky-Shapiro et al., 1996) (Han y Kamber, 2001). El KDD es básicamente un proceso automático en el que se combinan descubrimiento y análisis.

El proceso consiste en la preparación de los datos, la realización del proceso de minería y la interpretación de los resultados obtenidos. En la figura 3 se muestra este proceso.

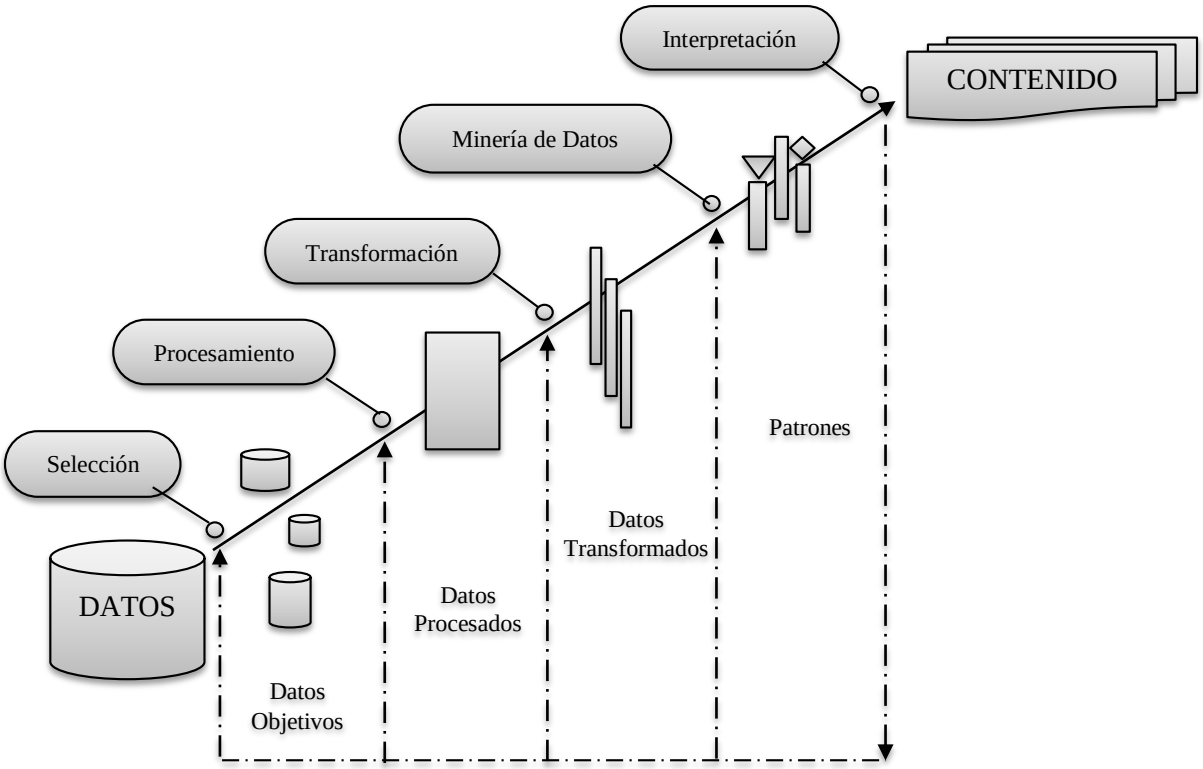

Figura 3 Etapas del proceso KDD Tomado de Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional (Timarán-Pereira, S. et al., 2016).

La minería de datos forma una parte central dentro de las etapas que se evidencian en KDD, para esto se tiene en cuenta una combinación de procesos como son: extracción de datos, limpieza, selección de características, algoritmos y análisis de resultados; en ello la metodología de análisis más ampliamente utilizada se conoce como CRISP-DM (Cross- Industry Standard Process for Data Mining).

2.2.1 Generalidades de Metodología CRISP-DM

Es uno de los modelos principalmente utilizados en los ambientes académicos e industriales y se constituye como la guía de referencia más ampliamente utilizada en el desarrollo de este tipo de proyectos (Hernández, Ramírez & Ferri, 2005).

Según Chapman et al., (2000) el ciclo de vida de CRISP-DM (ver figura 4) es dinámico e iterativo, por lo que la ejecución de los procesos no es estricta y con frecuencia se puede pasar de un proceso a otro, de atrás hacia adelante y viceversa. Esta metodología contempla seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación, como se muestra en la figura 4, y se describen a continuación.

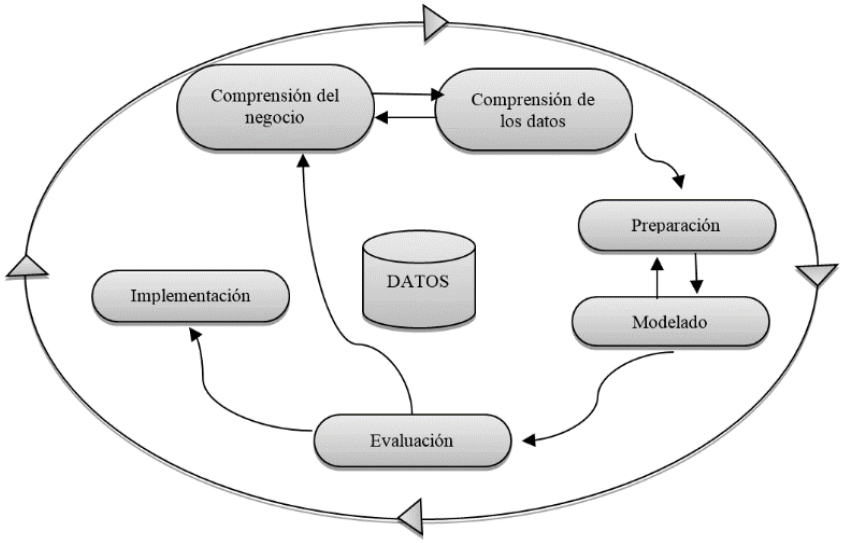

Figura 4 Ciclo de vida de CRISP-DM. Tomado de CRISP-DM 1.0 Step by-Step Data Mining Guide (P. Chapman Et al., 2000).

2.2.1.1 Comprensión del Negocio o Problema

Se centra en entender los objetivos y requerimientos desde un punto de vista empresarial o institucional para convertirlos en objetivos técnicos o en un plan de proyectos. En esta fase, se realizarán las actividades que permitan profundizar y apropiar de una manera completa el problema objeto de estudio, los objetivos y los requisitos de esta investigación, que posibiliten la recolección de los datos correctos para interpretar adecuadamente los resultados (Chapman et al., 2000). Esta fase abarca las siguientes sub-tareas:

La determinación de los Objetivos: permite establecer cuál es el negocio que se quiere resolver y se determina los criterios de éxito que pueden ser cuantitativos y cualitativos.

Evaluar la Situación Actual: se parte de los conocimientos previos acerca del negocio y se analiza las ventajas que trae aplicar la minería de datos.

Determinar los Objetivos de la Minería de Datos: se presenta los objetivos del negocio en términos de metas del proyecto.

Producir un Plan de Proyecto: se considera los pasos que se deben tener en cuenta para el desarrollo del proyecto y las metodologías a emplear para cada paso.

2.2.1.2 Comprensión de los Datos

En esta fase, se debe identificar y recopilar la información con el fin de familiarizarse con la información disponible en las bases de datos (Chapman et al., 2000). Los subprocesos que se siguen para esta fase son:

Recolectar los Datos Iniciales: se debe construir un repositorio inicial de datos realizar un informe con los datos adquiridos, describiendo su localización, las técnicas usadas para la recolección y los problemas asociados y las soluciones propias del proceso.

Describir los Datos: se debe describir cada atributo a través de un diccionario de datos, realizar un análisis de éstos y determinar su calidad con el fin de asegurar su completitud y corrección.

Explorar los Datos: es importante hacer un trabajo de exploración de los datos del repositorio a través de un análisis preliminar, con el fin de determinar qué variables son potencialmente importantes para el estudio para esto se pueden aplicar técnicas estadísticas.

Verificar la Calidad de los Datos: se debe determinar la consistencia de los valores en cada campo, la cantidad de datos nulos y los valores con algún tipo de ruido, con el fin de obtener completitud y corrección en los datos.

2.2.1.3 Preparación de los Datos

Se trata de preparar los datos para la posterior aplicación de técnicas de Minería de Datos, buscando patrones o relaciones entre las variables. Como resultado de esta fase se obtendrá un repositorio de datos limpio y transformado, listo para aplicarle las técnicas de Minería de Datos (Chapman et al., 2000). Los pasos a seguir en esta fase son:

Seleccionar los Datos: se seleccionarán del repositorio inicial construido, los atributos más representativos que permitan descubrir información relevante para el estudio.

Limpiar los Datos: se deben limpiar e integrar los datos en caso necesario, además; se han de generar atributos adicionales a partir de los existentes, realizar transformaciones o cambios de formato a los valores de los atributos necesarios.

Estructurar los Datos: aquí se generan nuevos atributos a partir de los ya conocidos, se integran nuevos registros o se transforma valores de atributos existentes.

Integrar los Datos: se crea nuevos campos y/o registros, se organiza nuevas tablas a partir de las ya existentes.

Formatear los Datos: se basa en la modificación de los datos sin pérdida de generalidad con el fin de aplicar alguna técnica de minería de datos como puede ser; eliminar comas, tabuladores, caracteres especiales, espacios, máximos y mínimos para las cadenas de caracteres, etc.

2.2.1.4 Modelado

Para la presente investigación se seleccionan las técnicas descriptivas de modelado más apropiadas para el proyecto de Minería de Datos y se aplican al repositorio limpio y transformado, resultado de la fase anterior, con el fin de obtener un modelo. Las técnicas a implementar deben cumplir con los siguientes criterios (Chapman et al., 2000):

- ✓ Ser apropiadas para el problema.
- ✓ Disponer de datos adecuados.
- ✓ Cumplir con los requisitos del problema.
- ✓ Técnicas adecuadas para obtener un modelo (criterio del investigador).

Para completar esta fase se incluyen los siguientes aspectos:

Generar Plan de Prueba: es necesario generar un plan de prueba que permita verificar la calidad y validez del modelo construido. Para esto se deben manejar dos conjuntos de datos; uno para entrenamiento y otro para prueba.

Construir el Modelo: se aplica la técnica seleccionada sobre los datos preparados para generar uno o más modelos. Las técnicas que usan en esta fase presentan un conjunto de parámetros, escogidos de forma iterativa, que determinan características del modelo a generar.

Evaluar el Modelo: En esta fase se evaluarán los patrones descubiertos con el fin de determinar su validez, remover los patrones redundantes o irrelevantes y traducir los patrones útiles en términos que sean entendibles para el usuario. Además, se revisará todo el proceso, con el fin de repetir un paso en el evento que se haya cometido un error (Chapman et al., 2000).

2.2.1.5 Evaluación

Una vez que el modelo ha sido construido y validado, el conocimiento obtenido se transforma en acciones dentro del proceso de negocio. En esta fase se hará un análisis y las posibles recomendaciones con respecto a los resultados obtenidos. Este conocimiento descubierto se podrá incorporar e integrar al existente para apoyar los procesos de toma de decisiones de acuerdo a los criterios de éxitos del problema (Chapman et al., 2000).

2.2.2 Modelos Predictivos y Modelos Descriptivos

Como se mencionó anteriormente la Minería de Datos se encarga de analizar datos con el fin de obtener conocimiento. Este conocimiento puede darse en forma de reglas, relaciones o patrones derivados de los datos y de antemano desconocidos, o también en forma resumida presentando información concisa y es precisamente esta la que constituyen el modelo de los datos, objeto de estudio. Para representar un modelo existen variadas formas y cada una de ellas determina el tipo de técnica a utilizar para inferir sobre los datos. En general pueden darse dos tipos de modelos: los predictivos y los descriptivos (Hernández et al, 2004).

2.2.2.1 Tareas Predictivas o Supervisadas

Según Hernández et al (2004) los modelos predictivos toman en cuenta variables objetivo o dependientes que son aquellas que se desea estimar y las variables predictivas o independientes que se usan para predecir los valores de interés en las variables dependientes. Un ejemplo clásico de este tipo de modelo es predecir el comportamiento esperado de un cliente a partir de los datos del perfil del usuario. Entre las técnicas que se destacan en los modelos predictivos están: la clasificación y la regresión.

2.2.2.1.1 Clasificación

Consiste en identificar clases o categorías las cuales se deben definir previamente para examinar las variables predictoras. Así se puede catalogar cada nuevo elemento en una clase definida. Esta tarea es una de las más frecuentes de la Minería de Datos. Ejemplos para este tipo de tarea son clasificar un mensaje de correo electrónico como spam o no, clasificar entre varios medicamentos cual es el mejor para una determinada patología (Riquelme, Ruiz & Gilbert, 2006).

2.2.2.1.2 Regresión

Consiste en examinar las variables predictoras para realizar una predicción numérica o regresión, es decir se busca encontrar similitudes entre los valores de los atributos de una

determinada clase de un conjunto dado. Esta terea se reconoce con otros nombre dependiendo de las características de los datos: interpolación (cuando el valor predicho está en medio de otros) o estimación (cuando se trata de algo futuro) por ejemplo: estimar ventas en el año 2016 en de una empresa multinacional, predecir el número de unidades defectuosas de un lote de productos, predecir la precisión de una válvula a partir de la entradas (Riquelme, Ruiz & Gilbert, 2006).

Algunas de las técnicas más comúnmente usadas en las tareas predictivas son:

Árboles de Decisión, describe las condiciones a causa de las decisiones tomadas (Breiman, 1984). Algoritmos de aplicación: C4.4, CLS, ID3, See5/C5.0.

Redes Neuronales, imita el funcionamiento interno de las neuronas humanas. Algoritmos de aprendizaje: Perceptrón, redes de base radial, hopfield.

Máquinas de Soporte Vectorial: Predice la clase correspondiente a partir de un hiperplano en alta dimensión que separa los elementos de las clases según su proximidad (Jiménez y Rengifo, 2010). Algoritmos: SVM, SVM light, Multiclass SVM, transductive SVM.

Métodos de Regresión: Permiten realizar predicciones numéricas por medio de funciones de regresión. Algoritmos: Regresión lineal, regresión logística.

2.2.2.2 Tareas Descriptivas o No Supervisadas

Estas tareas buscan identificar patrones que caracterizan o resumen los datos, por tanto sirven para explorar las propiedades de los datos examinados y no para hacer predicción de nuevos datos (Hernández et al., 2004). Las técnicas más comúnmente usadas en este modelo son: el análisis correlacional y las reglas de asociación.

2.2.2.2.1 Análisis Correlacional

Constituye una técnica estadística que nos indica si dos variables están relacionadas o no. La correlación puede decir algo acerca de la relación entre las variables. Se utiliza para entender

si la relación es positiva o negativa y la fuerza de la relación. Dichas correlaciones son medidas por el coeficiente de correlación (r), donde su valor numérico varía de 1,0 a $-1,0$. En general, $r > 0$ indica una relación positiva y $r < 0$ indica una relación negativa, mientras que $r = 0$ indica que no hay relación. Por ejemplo se puede analizar en un centro de salud los factores que influyen para que un paciente pueda asistir a dicho establecimiento, obtener información para la prevención de incendios, donde se desea conocer las correlaciones negativas entre el empleo de distintos grosores de protección del material eléctrico y la frecuencia con que ocurren los incendios (Hernández et al, 2004).

2.2.2.2.2 Agrupamiento (Clustering)

El proceso de agrupar objetos físicos o abstractos en clases de objetos similares se llama segmentación o clustering o clasificación no supervisada (M.-S. Chen et al., 1996). Básicamente, el clustering agrupa un conjunto de datos (sin un atributo de clase predefinido) basado en el principio de: maximizar la similitud intraclase y minimizar la similitud interclase. El análisis de clustering ayuda a construir particiones significativas de un gran conjunto de objetos basado en la metodología "divide y conquista" la cual descompone un sistema de gran escala en pequeños componentes para simplificar el diseño y la implementación. Uno de los algoritmos empleado para el análisis de esta técnica es k-means (Huang, 1997) (Huang, 1998) y tiene como propósito buscar similitudes entre los registros por medio de medidas de distancia, de tal manera que los elementos en el mismo clúster están cercanos entre sí (Moody y Darken, 1989).

El algoritmo tiene una fase de entrenamiento, que puede ser lenta, dependiendo del número de puntos a clasificar y de la dimensión del problema. Una vez terminado el entrenamiento, la clasificación de nuevos datos es muy rápida, ya que la comparación de las distancias se realiza solo con los prototipos.

El procedimiento es el siguiente:

Se calcula, para cada ejemplo x_k, el prototipo más próximo A_g y se incluye en la lista de ejemplos de dicho prototipo.

$$A_g = \arg_{\mathrm{A_i}}\min\{d(x_k, A_i)\} \qquad \forall_i = 1, \dots, n$$

Después de haber introducido los ejemplos, cada prototipo A_k tendrá un conjunto de ejemplos a los que representa:

$$l(A_k) = \{x_{k_1}, x_{k_2}, \dots, x_{k_m}\}$$

Se desplaza el prototipo hacia el centro de masas de su conjunto de ejemplos.

$$A_k = \frac{\sum_{i=1}^{m} x_{k_i}}{m}$$

Se repite el procedimiento hasta que ya no se desplazan los prototipos.

Mediante este algoritmo el espacio de ejemplos de entrada se divide en k clases o regiones, y el prototipo de cada clase estará en el centro de la misma. Dichos centros se determinan con el objetivo de minimizar las distancias cuadráticas euclídeas entre los patrones de entrada y el centro más cercano, es decir minimizando el valor J.

$$J = \sum_{i=1}^{k} \sum_{n=1}^{m} M_{i,n} d_{UECL}(x_n - A_i)^2$$

Donde m es el conjunto de patrones, d_{UECL} es la distancia euclídea, x_n es el ejemplo de entrada de n, A_i es el prototipo de la clase i, y $M_{i,n}$ es la función de pertenencia del ejemplo n a la región i de forma que vale 1 si el prototipo A_i es el más cercano al ejemplo x_n y 0 en caso contrario, es decir:

$$M_{i,n} = \begin{cases} 1 \text{ si } d_{UECL}(x_n - A_i) < d_{UECL}(x_n - A_s) \ \ \forall_s \neq i, s = 1,2, \dots, k \\ 0 \text{ en caso contrario} \end{cases}$$

La colocación final de los prototipos definirá la solución de entrada por el algoritmo. En la figura 5 se muestra un ejemplo de este proceso (Hernández J., et al., 2005).

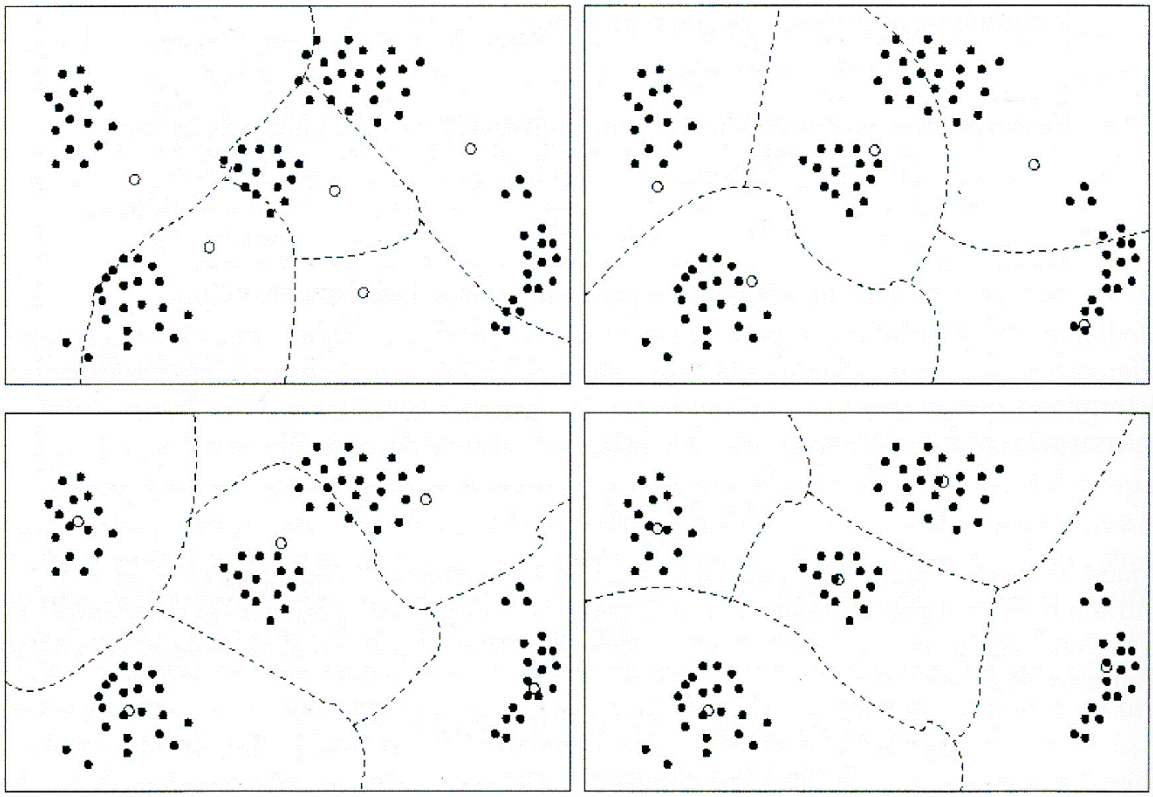

Figura 5 Ejemplo de evolución de los prototipos y grupos formados con el algoritmo k-means

2.2.2.2.3 Reglas de Asociación

Tiene como objetivo identificar relaciones no explicitas entre atributos categóricos. Estas reglas no implican una relación causa efecto, es decir, puede no existir una causa para que los datos estén asociados. Para este modelo se usan dos funciones de puntuación definidas como el soporte y la confianza las cuales son valores numéricos. El modelo matemático fue propuesto en (Agrawal et al, 1993) para afrontar el problema de minar reglas de asociación.

- ✓ **Soporte y Confianza**

La confianza es una medida de la efectividad de la regla. Representa el porcentaje de casos en los que dada la premisa se verifica la implicación. Así, para una asociación $A \rightarrow C$ (Reyes, J. & García, R. 2005):

$$sop(A \rightarrow C) = sop(A \cup C)$$

$$conf(A \rightarrow C) = \frac{sop(A \cap C)}{sop(A)}$$

La confianza denota la "fuerza" de la implicación y el soporte indica la frecuencia de ocurrencia de los patrones en la regla. Es a menudo deseable poner atención solamente a esas reglas que pueden tener razonablemente un soporte alto, ya que. Tales reglas con una confianza y soporte altos son referidas como reglas fuertes (strong rules) en (Agrawal et al, 1993). En (Agrawal et al, 1993), (Agrawal, Srikant, 1994), el problema de minar reglas de asociación es encontrar todas las reglas de asociación que satisfagan un soporte mínimo especificado por el usuario y una mínima restricción de confianza. El método se descompone en los siguientes pasos:

1. Descubrir los itemsets frecuentes, i.e., el conjunto de itemsets que tienen el soporte de transacciones por encima de un predeterminado soporte s mínimo.

2. Usar los itemsets frecuentes para generar las reglas de asociación para la base de datos.

✓ **Método para la Generación de Itemsets Candidatos**

Los algoritmos que generan itemsets candidatos para el descubrimiento de itemsets frecuentes hacen múltiples iteraciones sobre los datos. Para ello se tiene en cuenta los siguientes pasos:

- En el primer paso, se calcula el soporte de los items individuales y se determina si son o no frecuentes, i.e., que cumplan el soporte mínimo.
- En los siguientes pasos, se parte con un conjunto semilla (Agrawal, Srikant, 1994) de itemsets que fueron frecuentes en el paso previo.

- Se usa este conjunto semilla para generar nuevos itemsets frecuentes, llamados itemsets candidatos, calculando el soporte actual para ellos durante este paso.
- Al final del paso, se determina cuáles de los itemsets candidatos son actualmente frecuentes y ellos pasan a ser la semilla para el próximo paso.
- Este proceso continua hasta no encontrar nuevos itemsets frecuentes.
- Se asume que los items en cada transacción están ordenados lexicográficamente.

El número de items en un itemset se denomina su tamaño. Un itemset de tamaño k es un k-itemset. Se usa la notación $C_1 * C_2 * \ldots * C_k$ para representar un k-itemset C compuesto de items $C_1, C_2, \ldots, C_k$, donde $C_1 < C_2 < \ldots < C_k$. Asociado con cada itemset esta un contador que almacena el soporte de este itemset. El contador es inicializado en 0 cuando el itemset es creado por primera vez.

En la tabla 3 se resume la notación usada por los algoritmos Apriori y sus variantes (Agrawal, Srikant, 1994).

Tabla 3 Notación del algoritmo Apriori

k-itemset	**Un itemset que tiene k items.**
L_k	Conjunto de k-itemsets frecuentes (que cumplen con el soporte mínimo). Cada elemento de este conjunto es una pareja (itemset, soporte).
C_k	Conjunto de k-itemsets candidatos (potencialmente itemsets frecuentes). Cada elemento de este conjunto es una pareja (itemset, soporte).
$\overline{C_k}$	Conjunto de k-itemsets candidatos asociados con los TIDS de las transacciones generadas.

Fuente: Agrawal (1994). Fast Algorithms for Mining Association Rules.

- ✓ **Generación de Itemsets Candidatos con la Función Apriori_gen()**

La función apriori_gen() toma como argumento L_{k-1}, el conjunto de todos los $(k-1)$ itemsets frecuentes y retorna un superconjunto del conjunto de todos los k-itemsets frecuentes. La función primero, en el paso del join, se junta L_{k-1} con L_{k-1}:

$L_1 = \{1 - \text{itemsets frecuentes }\};$

For $(k = 2;\ L_{k-1} \neq \emptyset; k++)$ **do begin**

$C_k = \text{apriori\_gen}(L_{k-1});//$ nuevos candidatos

Forrall transacciones $t \in D$ **do begin**

$C_t = s\text{ubconjunto}(C_{k,t});//\text{candidatos contenidos en t}$

Forrall candidatos $c \in C_t$ **do**

$\text{c.count}++;$

end

$L_k = \{c \in C_k | \text{c.count} \geq \text{minsup}\}$

End

Respuesta

$\bigcup_k L_k$

Figura 6 Algoritmo A priori (Fuente: Agrawal, Srikant, 1994)

Insert into C_k

Select p.item_1, p.item_2,..., p.item_{k-1}, q.item_{k-1}

from $L_{k-1}p$, $L_{k-1}q$

where $p.item_1 = q.item_1$,..., $p.item_{k-2} = q.item_{k-2}$, $p.item_{k-1} < q.item_{k-1}$;

Luego, en el paso de poda, se borran todos los itemsets $c \in C_k$ tal que algún $(k-1)$ subconjunto de c no está en L_{k-1}:

Forall itemsets $c \in C_k$ do

Forall $(k-1)$ subconjuntos s de c do

If (s $\notin L_{k-1}$ **then**

Delete c from C_k;

- ✓ **Generación de Reglas**

Por cada itemset frecuente I, se generan todas las reglas $a \implies (I - a)$, donde a es un subconjunto de I, tal que

$$\frac{sop(I)}{sop(a)} \geq mincof$$

Para todo $a' \subseteq a$, $sop(a) \geq sop(a')$

Por consiguiente, la confianza de la regla $a' \implies (I - a')$ no puede ser mayor ($\leq$) que la confianza de $a \implies (I - a)$.

Si la regla $(I - a) \implies a$ se cumple, todas las reglas de la forma $(I - a') \implies a'$ deben cumplirse donde $a' \neq \emptyset \subseteq a$. Por ejemplo, si la regla $AB \implies CD$ se satisface, entonces las reglas $ABC \implies D$ y $ABD \implies C$ también se deben satisfacer.

De un itemset frecuente I, primero se generan todas las reglas con un item en el consecuente. Luego se usa los consecuentes de estas reglas y la función apriori_gen() para generar todos los posibles consecuentes con dos items que puedan aparecer en una regla generada a partir de l, etc. El algoritmo que usa esta idea se puede mirar en (Agrawal et al., 1996).

3. Materiales y Métodos

La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. Se siguieron las diferentes etapas de la metodología CRISP – DM con el fin de detectar patrones de desempeño académico en las Pruebas Saber 5° de los estudiantes pertenecientes a las instituciones educativas del país.

3.1. Comprensión del Negocio o Problema

3.1.1. Contexto

El problema de investigación se basa en el desempeño académico de los estudiantes de las instituciones educativas colombianas en las Pruebas Saber 5° del cual se pretende encontrar patrones asociados al buen o mal rendimiento de las mismas, en aspectos socioeconómicos, académicos e institucionales. Para esto, en primera instancia se cuenta con la base de datos proporcionada por el ICFES de los años 2014 al 2016 en donde se evidencian los elementos mencionados anteriormente. Además, se profundizó en aspectos teóricos sobre rendimiento escolar, lineamientos de Pruebas Saber, factores asociados al desempeño académico, competencias genéricas, alineación y estructura de las Pruebas Saber. Tal como se definió en el segundo capítulo de esta investigación.

3.1.2. Objetivo

Descubrir factores asociados al desempeño académico en las competencias que evalúan las Pruebas Saber 5° de los estudiantes pertenecientes a instituciones educativas colombianas, a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES en el periodo 2014 a 2016, a través de técnicas descriptivas de Minería de Datos, que

permitan generar conocimiento encaminado a soportar las decisiones institucionales y gubernamentales para el mejoramiento de la calidad educativa.

3.2. Comprensión de los Datos

En esta fase se hizo la identificación, recopilación y familiarización con la base de datos proporcionada por el ICFES, sobre el rendimiento de las Pruebas Saber 5 en los años 2014 al 2016, y con los datos socioeconómicos, académicos e institucionales, de los estudiantes que presentaron esta prueba a nivel nacional.

3.2.1 Descripción de Diccionario de Datos Inicial

El diccionario inicial contiene la descripción de cada una de las variables de las diferentes tablas como son: valores plausibles, instituciones, sedes, municipios, departamentos, entidades territoriales, entre otras. El repositorio inicial contenía tres bases de datos que se nombraron según el año de presentación de las pruebas saber 5 como: saber5_2014, saber5_2015 y saber5_2016; estas tres bases de datos se unificaron en un solo repositorio llamado saber5_2014_2016 el cual cuenta con 2’038.120 registros y 47 atributos. En la tabla 4 se muestra el diccionario de datos de los valores plausibles (en adelante estudiantes), que incluye el nombre del campo, la descripción de las variables y el tipo de campo.

Tabla 4 Diccionario de datos de valores plausibles (estudiantes).

Nombre de Campo	Descripción	Tipo de Campo
ID_hoja	Código de identificación de hoja de respuestas	VAAAAGXXXXXXX donde V = constante, AAAA = Año, G = grado y XXXXXXX = consecutivo

Grupo	Grupo/salón al que pertenecen los estudiantes	99 = censal , ¬99 = control
N	Matrícula del grupo al que pertenecen	1, 2, …,
Estrato	Estrato al que pertenece el establecimiento dentro del marco muestral	Alfanumérico(9)
Aplicación	No definido	X
Grado	Grado del estudiante	3 , 5 , 9
Jornada	(Codsitio) - Código de la sede-jornada al que pertenece el estudiante()	Numérico(6) Ir a tabla sedes en la columna ID
Establecimiento	Código DANE del establecimiento educativo al que pertenece el estudiante	Llave a tabla INSTITUCION
EnteTerr	Código DANE de la entidad a la que pertenece establecimiento educativo	Llave a tabla ENTIDADTERRITORIAL
Departamento	Código DANE del departamento al que pertenece establecimiento educativo	Llave a tabla DEPARTAMENTO
Municipio	Código DANE del municipio al que pertenece establecimiento educativo	Llave a tabla MUNICIPIO
Género	Sexo del estudiante	1=Masculino , 2=Femenino

		, 3= No especifica
Sector	Sector del establecimiento	1 = Oficial , 2 = No oficial
Zona	Zona donde se ubica la mayoría de la población atendida	1 = Urbana , 2 = Rural
ZonaStab	Zona del establecimiento	1 = Urbana , 2 = Rural
Calendario	Calendario del establecimiento	A , B
Nivel Socioeconómico	Nivel socioeconómico del establecimiento	1 , 2 , 3 , 4 , 5
ModeloEdu	No definido	X
Disenso	Marca de discapacidad cognitiva	0 = sin discapacidad, 1 = con discapacidad
Leng_copietas	Indicador de copia de lenguaje	0 = no copia, 1 = con copia
Leng_weight	Peso muestral de lenguaje	Numérico(6)
Leng_score1	Score1 - Valor plausible 1 de lenguaje	Numérico(6)
Leng_score2	Score2 - Valor plausible 2 de lenguaje	Numérico(6)
Leng_score3	Score3 - Valor plausible 3 de lenguaje	Numérico(6)
Leng_score4	Score4 - Valor plausible 4 de lenguaje	Numérico(6)
Leng_score5	Score5 - Valor plausible 5 de lenguaje	Numérico(6)

Mate_copietas	Indicador de copia de matemáticas	0 = no copia, 1 = con copia
Mate_weight	Peso muestral de matemáticas	Numérico(6)
Mate_score1	Score1 - Valor plausible 1 de matemáticas	Numérico(6)
Mate_score2	Score2 - Valor plausible 2 de matemáticas	Numérico(6)
Mate_score3	Score3 - Valor plausible 3 de matemáticas	Numérico(6)
Mate_score4	Score4 - Valor plausible 4 de matemáticas	Numérico(6)
Mate_score5	Score5 - Valor plausible 5 de matemáticas	Numérico(6)
Cien_copietas	Indicador de copia de ciencias	0 = no copia, 1 = con copia
Cien_weight	Peso muestral de ciencias	Numérico(6)
Cien_score1	Score1 - Valor plausible 1 de ciencias	Numérico(6)
Cien_score2	Score2 - Valor plausible 2 de ciencias	Numérico(6)
Cien_score3	Score3 - Valor plausible 3 de ciencias	Numérico(6)
Cien_score4	Score4 - Valor plausible 4 de ciencias	Numérico(6)
Cien_score5	Score5 - Valor plausible 5 de	Numérico(6)

	ciencias	
Comp_copietas	Indicador de copia de competencias ciudadanas	0 = no copia, 1 = con copia
Comp_weight	Peso muestral de competencias ciudadanas	Numérico(6)
Comp_score1	Score1 - Valor plausible 1 de competencias ciudadanas	Numérico(6)
Comp_score2	Score2 - Valor plausible 2 de competencias ciudadanas	Numérico(6)
Comp_score3	Score3 - Valor plausible 3 de competencias ciudadanas	Numérico(6)
Comp_score4	Score4 - Valor plausible 4 de competencias ciudadanas	Numérico(6)
Comp_score5	Score5 - Valor plausible 5 de competencias ciudadanas	Numérico(6)

Fuente. Tomado de las bases de datos ICFES.

Además se cuenta con tablas adicionales cuyos resultados vienen dados por establecimiento educativo y se muestran en dos tipos diferentes de reporte, dependiendo de la cantidad de estudiantes que presentaron la prueba: los reportes completos y los simplificados.

En la tabla 5 se indica los reportes completos que son usados para indicar las estadísticas de cada establecimiento en los que por cada grado_área participaron seis o más estudiantes en la prueba. Las variables contenidas en estos reportes son el promedio de calificación de los

estudiantes, el error estándar del promedio y la desviación. Además, las variables de nivel de desempeño contienen el porcentaje de estudiantes participantes que se ubican en cada nivel.

Tabla 5 Resultados instituciones completo.

Nombre de Campo	Tipo de Campo	Descripción
Cod_Dane	Texto	Código DANE del establecimiento educativo
Evaluados	Numérica	Total de estudiantes del establecimiento educativo que participan en la evaluación.
Participantes	Numérica	Total de estudiantes del establecimiento educativo que fueron
Copia	Texto	Indicio de copia en el área: 0=No se presentan indicios de copia, 1=Indicios de copia individual, 2=Indicios de copia masiva en alguna sede-jornada
Peso	Numérica	Peso o ponderación a usar para el cálculo de agregados con la información a nivel de sede jornada
Promedio	Numérica	Promedio de los puntajes de los estudiantes dentro del Establecimiento Educativo
Error_Estandar	Numérica	El error estándar del promedio de los puntajes de los estudiantes dentro del Establecimiento Educativo
Desviación	Numérica	Desviación estándar del puntaje de los estudiantes dentro del Establecimiento Educativo

Insuficiente	Numérica	Porcentaje de estudiantes en el nivel de desempeño insuficiente
Mínimo	Numérica	Porcentaje de estudiantes en el nivel de desempeño mínimo
Satisfactorio	Numérica	Porcentaje de estudiantes en el nivel de desempeño satisfactorio
Avanzado	Numérica	Porcentaje de estudiantes en el nivel de desempeño avanzado

Fuente. Tomado de las bases de datos ICFES.

En la tabla 6 se encuentra la descripción de variables de los reportes simplificados que pertenecen a los establecimientos educativos en los que participaron menos de seis estudiantes por grado_área. En este reporte no está contenido el promedio, el error y la desviación estándar; a diferencia del reporte completo. En las variables de nivel de desempeño se encuentra el número, más no el porcentaje de estudiantes que se ubican en cada nivel.

Tabla 6 Resultados Instituciones_simplificado.

Nombre de Campo	Tipo de Campo	Descripción
Cod_Dane	Texto	Código DANE del establecimiento educativo
Evaluados	Numérica	Total de estudiantes del establecimiento educativo que participan en la evaluación.
Participantes	Numérica	Total de estudiantes del establecimiento educativo que fueron evaluados en el área.

Copia	Texto	Indicio de copia en el área: 0=No se presentan indicios de copia, 1=Indicios de copia individual, 2=Indicios de copia masiva en alguna sede-jornada
Peso	Numérica	Peso o ponderación a usar para el cálculo de agregados con la información a nivel de sede-jornada
Insuficiente	Numérica	Número de estudiantes en el nivel de desempeño insuficiente
Mínimo	Numérica	Número de estudiantes en el nivel de desempeño mínimo
Satisfactorio	Numérica	Número de estudiantes en el nivel de desempeño satisfactorio
Avanzado	Numérica	Número de estudiantes en el nivel de desempeño avanzado

Fuente. Tomado de las bases de datos ICFES.

La tabla 7 muestra los resultados simplificados a nivel de agregación de la sede_jornada, e indican el número de estudiantes por cada nivel de desempeño.

Tabla 7 Resultados Sede_jornada.

Nombre de Campo	Tipo de Campo	Descripción
Id_Sede	Texto	Código saber de la sede jornada
Codigo_Dane_Sede	Texto	Código DANE de la sede del establecimiento educativo
Jornada	Texto	Jornada en la que desempeña actividades y para la que se reporta resultados. M=Mañana, T=Tarde, C=Completa
Evaluados	Numérica	Total de estudiantes del establecimiento educativo que

		participan en la evaluación.
Participantes	Numérica	Total de estudiantes del establecimiento educativo que fueron evaluados en el área
Copia	Texto	Indicio de copia en el área: 0=No se presentan indicios de copia, 1=Indicios de copia individual, 2=Indicios de copia masiva en alguna sede-jornada
Insuficiente	Numérica	Número de estudiantes en el nivel de desempeño insuficiente
Mínimo	Numérica	Número de estudiantes en el nivel de desempeño mínimo
Satisfactorio	Numérica	Número de estudiantes en el nivel de desempeño satisfactorio
Avanzado	Numérica	Número de estudiantes en el nivel de desempeño avanzado

Fuente. Tomado de las bases de datos ICFES.

La tabla 8 muestra los resultados simplificados a nivel de agregación de municipio, e indican el número de estudiantes por cada nivel de desempeño.

Tabla 8 Resultados Municipio

Nombre de Campo	Tipo de Campo	Descripción
Muni_Id	Texto	Código saber del municipio
Municipio	Texto	Nombre del municipio según divipo

Departamento	Texto	Nombre del departamento
Puntaje_Promedio	Numérica	Promedio de los puntajes de los estudiantes dentro del Establecimiento Educativo
Errorestandar Promedio	Numérica	El error estándar del promedio de los puntajes de los estudiantes dentro del Establecimiento Educativo
Desviación	Numérica	Desviación estándar del puntaje de los estudiantes dentro del Establecimiento Educativo
Insuficiente	Numérica	Número de estudiantes en el nivel de desempeño insuficiente
Mínimo	Numérica	Número de estudiantes en el nivel de desempeño mínimo
Satisfactorio	Numérica	Número de estudiantes en el nivel de desempeño satisfactorio
Avanzado	Numérica	Número de estudiantes en el nivel de desempeño avanzado
N	Numérica	Número de participantes

Fuente. Tomado de las bases de datos ICFES.

A continuación se muestran tablas adicionales de identificación (tabla 9 hasta tabla 13) que contienen información sobre entidad territorial, municipio, departamentos, establecimiento educativo y sedes.

Tabla 9 Identificación del campo Entidades.

Nombre de Campo	Tipo de Campo	Descripción
Id_Ente	Texto	Identificador del ente territorial
Nombre	Texto	Nombre del ente territorial
Munexclu	Texto	Nombre de los municipios certificados excluidos dentro del ente Territorial
Tipo	Texto	Tipo de ente territorial 1=ETC,2=DPTO,3=MPIO

Fuente. Tomado de las bases de datos ICFES.

Tabla 10 Identificación del campo Municipios

Nombre de Campo	Tipo de Campo	Descripción
Id_Municipio	Texto	Código dane del municipio según divipo
Nombre	Texto	Nombre del municipio según divipo

Fuente. Tomado de las bases de datos ICFES.

Tabla 11 Identificación del campo Departamentos

Nombre de Campo	Tipo de Campo	Descripción
Id_Municipio	Texto	Código dane del municipio según divipo
Nombre	Texto	Nombre del municipio según divipo

Fuente. Tomado de las bases de datos ICFES.

Tabla 12 Identificación del campo Establecimientos

Nombre de Campo	Tipo de Campo	Descripción
Cod_Dane	Texto	Código DANE del establecimiento educativo
Id_Municipio	Texto	Código dane del municipio al que pertenece el establecimiento Educativo
Id_Ente	Texto	Id de la entidad territorial
Nombre	Texto	Nombre del establecimiento educativo reportado en el DUE
Zona	Texto	Zona donde la mayoría de la población atendida por el establecimiento educativo se ubica. 1=Urbano, 2=rural
Sector	Texto	Naturaleza administrativa del establecimiento educativo. 1= Oficial,2=No oficial
Tipo_Estab	Texto	Tipo de establecimiento. 1=Oficial urbano, 2=Oficial rural, 3=No Oficial
Calendario	Texto	Calendario del establecimiento
Nivel_Socio	Texto	NSE asignado de acuerdo a la clasificación realizada con puntajes Promedios

Fuente. Tomado de las bases de datos ICFES.

Tabla 13 Identificación del campo Sedes

Nombre de Campo	Tipo de Campo	Descripción
Id	Texto	Código saber de la sede jornada
Codigo_Dane_Estab	Texto	Código DANE del establecimiento educativo

Codigo_Dane_Sede	Texto	Código DANE de la sede del establecimiento educativo
Jornada	Texto	Jornada en la que desempeña actividades y para la que se reporta resultados. M=Mañana, T=Tarde, C=Completa
Nombre	Texto	Nombre de la sede jornada según DUE

Fuente. Tomado de las bases de datos ICFES.

Las tablas que se muestran a continuación (tabla 14 hasta tabla 22) fueron adicionadas por conveniencia ya que en ellas se encontraban datos implícitos como valores de 0 y 1, iniciales de palabras, entre otras y se renombró con caracteres.

Tabla 14 Descripción de Indicio de Copia

ID	Descripción
0	No se presentan indicios de copia
1	Indicios de copia individual
2	Indicios de copia masiva en alguna sede-jornada

Nota. Elaboración propia

Tabla 15 Descripción de la Jornada

ID	Descripción
M	Mañana
T	Tarde
C	Completa
U	Única

Nota. Elaboración propia

Tabla 16 Descripción del Tipo de Entidad

ID	Descripción
0	Entidad Territorial Certificada
1	Departamento
2	Municipio
4	Municipio si
5	No certificadas

Nota. Elaboración propia

Tabla 17 Descripción de la Zona

ID	Descripción
1	Urbano
2	Rural

Nota. Elaboración propia

Tabla 18 Descripción de Discapacidad

ID	Descripción
0	Sin Discapacidad
1	Con Discapacidad

Nota. Elaboración propia

Tabla 19 Descripción del Sector

ID	Descripción
1	Oficial
2	No Oficial

Nota. Elaboración propia

Tabla 20 Descripción del Tipo de establecimiento

ID	Descripción
1	Oficial urbano
2	Oficial rural
3	No Oficial

Nota. Elaboración propia

Tabla 21 Descripción del Género

ID	Descripción
1	Masculino
2	Femenino
3	No Especifica

Nota. Elaboración propia

Tabla 22 Descripción de Copietas

ID	Descripción
0	No copia
1	Con copia

Nota. Elaboración propia

Los resultados de la exploración de la base de datos, a través del análisis de tendencias de desempeño académico en las 4 competencias genéricas estudiadas, se muestran en el siguiente apartado.

3.2.2. Tendencias de Desempeño Académico en Competencias Genéricas–Pruebas Saber 5

Con el objetivo de explorar los datos se presenta en la tabla 4 un análisis de correlaciones entre las cuatro competencias genéricas de las pruebas Saber 5 (Matemáticas, Lenguaje, Ciencias y Competencias Ciudadanas) se estableció la tendencia estadística del desempeño académico en ellas con relación a aspectos socioeconómicos, académicos e institucionales. Para esto se utilizó la base de datos de las pruebas Saber 5 en los años 2014 – 2016 del ICFES y se seleccionaron los datos de valores plausibles; a través del coeficiente de correlación de Pearson, se establece cómo se asocian linealmente, entre sí, las cuatro competencias. Los resultados de presentan a continuación:

Tabla 23 Análisis de correlación entre 1as Competencias Genéricas

Competencias Genéricas	Coeficiente de Correlación	Lenguaje	Matemáticas	Ciencias Naturales	Competencias Ciudadanas
Lenguaje	Coef. de correlación	1	0,999**	0,999**	0,759**
	P valor		0,000	0,000	0,000
	N	1354409	675125	449726	219748
Matemática	Coef. de correlación	0,999**	1	0,999**	0,748**
	P valor	0,000		0,000	0,000
	N	675125	1352811	448351	220248
Ciencias Naturales	Coef. de correlación	0,999**	0,999**	1	.b
	P valor	,000	0,000		.
		449726	448351	902127	0

	N				
Competencias Ciudadanas	Coef. de correlación	0,759**	0,748**	.b	1
	P valor	0,000	0,000	.	
	N	219748	220248	0	441971

Nota. Elaboración propia

**. La correlación es significativa en el nivel 0,01 (2 colas).
b. No se puede calcular porque, como mínimo, una de las variables es constante.

De acuerdo a los resultados obtenidos en la tabla 23, se observa que Lenguaje con Matemáticas, Ciencias Naturales y Competencias Ciudadanas presentan correlaciones altas (r>0,5) por lo que se espera que los estudiantes que tienen buen desempeño en la prueba de Lenguaje también tengan buen desempeño en Matemáticas y Ciencias Naturales.

- ✓ **Desempeño en las Cuatro Competencias Genéricas Según Variables Socioeconómicas, Académicas e Institucionales**

Las tendencias de desempeño académico de los estudiantes y las cuatro competencias genéricas de las pruebas Saber 5, en relación con los aspectos socioeconómicos, académicos e institucionales se obtuvieron cruzando los puntajes de las competencias obtenidos en la prueba con las variables mencionadas anteriormente y se utilizaron las medias estadísticas: promedio, desviación estándar, intervalos de confianza y el tamaño del efecto de la diferencia estandarizada de las medias *d* de Cohen.

Para establecer si las diferencias en puntajes obtenidos en las cuatros competencias son estadísticamente significativas, se calcularon los intervalos de confianza de medias al 95%. Por otra parte, para determinar el tamaño de las diferencias de los promedios entre los diferentes grupos de variables analizadas se utilizó el estadístico d de Cohen con la siguiente escala propuesta por

Cohen (1998): [0.0, 0.2] trivial o muy pequeña, [0.5, 0.8] moderada y [0.8, infinito] grande. Este estadístico se obtiene dividiendo la diferencia de medias (en valor absoluto) de los dos grupos por comparar entre la desviación estándar conjunta de estos. Para su cálculo se utiliza como referencia el grupo del más alto promedio en cada competencia. A continuación se presenta el análisis de desempeño de los estudiantes en las pruebas Saber 5 en los años 2014 a 2016 según género, sector y tipo de establecimiento educativo, zona, nivel socioeconómico y jornada.

3.2.2.1. Género y Desempeño Académico en Competencias Genéricas

En la tabla 24 se muestra el desempeño en las competencias genéricas según el género. A partir de esta tabla se concluye que los hombres presentan mejor desempeño que las mujeres en tres de las competencias genéricas excepto en competencias ciudadanas en la cual sobresale el ítem donde no se especificó el género, sin embargo según el estadístico d de Cohen estas diferencias son de magnitud trivial en las cuatro competencias.

Tabla 24 Desempeño académico en competencias genéricas según género

Competencias Genéricas	Género	N	Media	Error Estándar	d
Lenguaje	Masculino	659534	2,63059	0,015254	
	Femenino	649656	2,58459	0,015065	0,004
	No Especifica	45219	2,22015	0,042783	0,034
Matemáticas	Masculino	657966	2,62053	0,015211	
	Femenino	647061	2,59965	0,015216	0,032
	No Especifica	47784	2,21759	0,041442	0,000
Ciencias Naturales	Masculino	438100	3,06901	0,022746	
	Femenino	431821	3,04039	0,022800	0,002
	No Especifica	32206	2,38292	0,058696	0,046

Competencias Ciudadanas	Masculino	216154	1,72101	0,000580	0,139
	Femenino	211634	1,72127	0,000578	0,140
	No Especifica	14183	1,75865	0,002491	

Nota. Elaboración propia

3.2.2.2. Sector y Desempeño Académico en Competencias Genéricas

En la tabla 25 se muestra el desempeño en las competencias genéricas según el sector del establecimiento educativo. A partir de esta tabla se concluye que en las pruebas de competencias genéricas Saber 5 del 2014 a 2016 los colegios no oficiales presentan mejor desempeño que los colegios oficiales, excepto en competencias ciudadanas en donde se observa una pequeña diferencia. Según el estadístico d de Cohen estas diferencias son de magnitud pequeña en las cuatro competencias.

Tabla 25 Desempeño académico en competencias genéricas según el sector del establecimiento

Competencias Genéricas	Sector	N	Media	Error Estándar	d
Lenguaje	Oficial	1084542	1,66780	0,000244	0,387
	No Oficial	269867	6,32036	0,051878	
Matemáticas	Oficial	1083401	1,66683	0,000243	0,387
	No Oficial	269410	6,33411	0,052000	
Ciencias Naturales	Oficial	717833	1,63893	0,000280	0,466
	No Oficial	184294	8,45228	0,075377	
Competencias Ciudadanas	Oficial	358132	1,73513	0,000464	
	No Oficial	83839	1,66770	0,000758	0,252

Nota. Elaboración propia

3.2.2.3. Zona y Desempeño Académico en Competencias Genéricas

A partir de la tabla 26 se puede concluir que los establecimientos educativos ubicados en zona urbana presentan mejor desempeño que los establecimientos ubicados en zona rural. Sin embargo, en competencias ciudadanas el promedio favorece a los establecimientos ubicados en zona rural. Según el estadístico d de Cohen estas diferencias son de magnitud trivial.

Tabla 26 Desempeño académico en competencias genéricas según la zona del establecimiento educativo

Competencias Genéricas	Zona	N	Media	Error estándar	D
Lenguaje	Urbano	1038212	2,70192	0,012675	
	Rural	316197	2,24316	0,016590	0,038
Matemáticas	Urbano	1037288	2,70404	0,012701	
	Rural	315523	2,24214	0,016593	0,038
Ciencias Naturales	Urbano	698897	3,17479	0,018794	
	Rural	203230	2,53573	0,025678	0,043
Competencias Ciudadanas	Urbano	332870	1,71820	0,000418	0,062
	Rural	109101	1,73498	0,001027	

Nota. Elaboración propia

3.2.2.4. Nivel Socioeconómico y Desempeño Académico en Competencias Genéricas

En la tabla 27 se puede observar que en las competencias de lenguaje, matemáticas y ciencias el desempeño académico es mayor en los establecimientos pertenecientes al estrato 4 y en competencia ciudadanas sobresale los establecimientos educativos pertenecientes al estrato 1. Según el estadístico d de Cohen estas diferencias son de magnitud pequeña en lenguaje, matemáticas y ciencias, y trivial en competencias ciudadanas.

Tabla 27 Desempeño académico en competencias genéricas según el nivel socioeconómico

Competencias Genéricas	Nivel Socioeconómico	N	Media	Error Estándar	D
Lenguaje	Estrato 1	270550	1,72348	0,002513	0,234
	Estrato 2	454009	1,76869	0,006411	0,262
	Estrato 3	359148	1,83246	0,009195	0,239
	Estrato 4	262696	5,99071	0,050666	
	Estrato 5	8006	1,66138	0,002933	0,169
Matemáticas	Estrato 1	269646	1,72104	0,002583	0,234
	Estrato 2	453242	1,76816	0,006410	0,263
	Estrato 3	359022	1,83249	0,009260	0,239
	Estrato 4	262765	5,99595	0,050671	
	Estrato 5	8136	1,64835	0,002760	0,170
Ciencias Naturales	Estrato 1	163264	1,71429	0,004114	0,318
	Estrato 2	343427	1,77552	0,008281	0,387
	Estrato 3	241566	1,87877	0,013369	0,338
	Estrato 4	150375	9,21202	0,087709	
	Estrato 5	3495	1,55649	0,002131	0,228

Competencias Ciudadanas	Estrato 1	103672	1,75605	0,001014	
	Estrato 2	105770	1,73995	0,000843	0,053
	Estrato 3	116516	1,71389	0,000722	0,147
	Estrato 4	111469	1,68232	0,000637	0,270
	Estrato 5	4544	1,74208	0,004649	0,043

Nota. Elaboración propia

3.2.2.5. Jornada y Desempeño Académico en Competencias Genéricas

En la tabla 28 se evidencia que el desempeño académico en las competencias genéricas de lenguaje, matemáticas y ciencias es mayor en los establecimientos que pertenecen a jornada de la mañana. En competencias ciudadanas se evidencia que tienen mayor rendimiento los establecimientos que pertenecen a la jornada completa. Según el estadístico d de Cohen las diferencias son de magnitud trivial en las cuatro competencias.

Tabla 28 Desempeño académico en competencias genéricas según jornada

Competencias Genéricas	Jornada	N	Media	Error Estándar	d
Lenguaje	Mañana	789551	2,12594	0,009574	0,207
	Tarde	3970	1,67961	0,003943	0,143
	Completa	247047	4,87333	0,045227	
	Única	313841	1,99243	0,013212	0,181
Matemáticas	Mañana	788035	2,12826	0,009657	0,206
	Tarde	3893	1,69697	0,003872	0,142
	Completa	247242			

	Única	313641	4,86837 1,99242	0,045147 0,013130	 0,181
Ciencias Naturales	Mañana Tarde Completa Única	509941 2164 174278 215744	2,34178 1,69131 6,20766 2,10667	0,014722 0,005652 0,063930 0,019189	0,238 0,170 0,215
Competencias Ciudadanas	Mañana Tarde Completa Única	273044 1709 71545 95673	1,72337 1,73669 1,68388 1,74790	0,000530 0,006752 0,001003 0,000779	0,092 0,046 0,253

Nota. Elaboración propia

3.2.1.6 Calendario Académico y Desempeño Académico en Competencias genéricas

Según la tabla 29 el desempeño académico en las competencias genéricas de Lenguaje, Matemáticas y Ciencias Naturales es mayor, puesto que pertenecen al calendario académico A. En competencias ciudadanas tienen mejor rendimiento los establecimientos que pertenecen al calendario B. Según el estadístico d de Cohen las diferencias entre las tres primeras competencias son de magnitud grande y en Competencias Ciudadanas son de magnitud trivial.

Tabla 29 Desempeño académico en competencias genéricas y calendario académico

Competencias Genéricas	Calendario	N	Media	Error Estándar	d
Lenguaje	A	1330214	1,654945	0,000211	5,270
	B	24195	4,26846	0,480035	
Matemáticas	A	1328561	1,65395	0,000211	5,258
	B	24250	54,22422	0,479382	
Ciencias Naturales	A	885381	1,62578	0,00024	6,994
	B	16746	77,31684	0,613718	
Competencias Ciudadanas	A	434570	1,72256	0,000409	
	B	7401	1,70963	0,002722	0,048

Nota. Elaboración propia

Después de realizar el análisis de las tendencias de desempeño académico en competencias genéricas y obtener algunas conclusiones relevantes, se procede a realizar la descripción de cada uno de los diccionarios de datos del repositorio inicial.

3.3 Preparación de los Datos

En esta fase los atributos que están contenidos en el repositorio de datos y que según el ICFES son los más importantes para capturar la información de las Pruebas Saber 5, fueron depurados teniendo en cuenta la calidad de los datos y las técnicas de Minería que se va aplicar

Inicialmente, se cuenta con tres repositorios correspondientes a los años 2014, 2015 y 2016; cada una de ellos contiene datos de las entidades territoriales, de los establecimientos educativos, las sedes adscritas a cada uno de ellos y a qué municipio pertenecen, además de esto, también se encuentran resultados por establecimientos educativos en dos denominaciones: completos (más de

seis estudiantes que presentan la prueba) y simplificado (menos de seis estudiantes que presentan la prueba), los últimos se eliminaron de la base ya que arrojaron sesgo en el resultado de análisis. Estos resultados se especifican por cada competencia genérica. Asimismo, se encuentran los resultados por municipio y por sedes. Por último, los repositorios contienen los valores plausibles que son los resultados de las pruebas individuales e incluyen datos como: la jornada, estrato, sexo, calendario, modelo educativo, municipio, zona, entre otros.

A continuación se describe los procesos realizados en las fases de limpieza y transformación de datos.

3.3.1 Limpieza

Uno de los requerimientos para aplicar las técnicas de minería de datos es que el repositorio esté limpio, es decir, que no exista presencia de datos faltantes o perdidos (missing values) o valores que no se ajusten al comportamiento general de los datos (outliers), para esto se ve la necesidad de mejorar la calidad de los datos contenidos en las bases de datos denominadas saber5_ 2014, saber5_2015, y saber5_2016 proporcionadas por el ICFES.

En primer lugar, se hizo un proceso de identificación y eliminación de valores de las variables que contenían caracteres no permitidos como tildes, diéresis, comillas, súper índices, subíndices, énfasis, símbolos numéricos como el porcentaje, fracciones, el vacío, exponenciales (Ø^Ä ÂÐ ½¾•Ý¢ªÇáéíóúÁÉÍÓÚäëïöüÄËÏÖÜÑñ') entre otros. Además, se cambiaron los datos de la tabla que aparecen vacíos como nulos, esto se hizo para cada una de las bases de datos mencionadas anteriormente.

Cada una de las bases de datos contiene una tabla de valores plausibles denominada estudiantes. Tomando como referencia la base saber5_2014 se creó una copia denominada *estudiantescopy* de la cual se identificaron datos de variables como sedes, establecimientos,

entidad territorial y municipios que no estaban o contenían valores erróneos, para luego buscarlos y completarlos de las bases de datos 2015 y 2016. Por ejemplo, en la variable establecimientos, se encontraron valores expresados en notación científica a los cuales se les aplica la actualización en forma de código. A los valores de las variables que no se logró completar con las bases de datos saber5_2015 y saber5_2016 se las eliminó. Esto se soporta con el porcentaje de datos nulos que se encontraban en las bases de datos como se especifican en la tabla 30.

Tabla 30 Atributos con un alto porcentaje de valores nulos

Año	Atributo	Nulos
2014	comp_copietas, com_weight, comp_score1, comp_score2, comp_score3, comp_score4 y comp_score5	100%
	cien_copietas, cien_weight, cien_score1, cien_score2, cien_score3, cien_score4 y cien_score5	34,06%
	mate_copietas, mate_weight, mate_score1, mate_score2, mate_score3, mate_score4 y mate_score5	33,72%
	leng_copietas, leng_weight, leng_score1, leng_score2, leng_score3, leng_score4 y leng_score5	33,61%
2015	cien_copietas, cien_weight, cien_score1, cien_score2, cien_score3, cien_score4 y cien_score5	100%

	comp_copietas, com_weight, comp_score1, comp_score2, comp_score3, comp_score4 y comp_score5	34,14%
	mate_copietas, mate_weight, mate_score1, mate_score2, mate_score3, mate_score4 y mate_score5	33,54%
	leng_copietas, leng_weight, leng_score1, leng_score2, leng_score3, leng_score4 y leng_score5	33,55%
2016	comp_copietas, com_weight, comp_score1, comp_score2, comp_score3, comp_score4 y comp_score5	100%
	cien_copietas, cien_weight, cien_score1, cien_score2, cien_score3, cien_score4 y cien_score5	34,12%
	mate_copietas, mate_weight, mate_score1, mate_score2, mate_score3, mate_score4 y mate_score5	33,76%
	leng_copietas, leng_weight, leng_score1, leng_score2, leng_score3, leng_score4 y leng_score5	33,52%

Nota. Elaboración propia

Los porcentajes del 100% de valores nulos se deben a que no se aplicó la competencia de ciencias en el año 2015 y la competencia de ciudadanía en los años 2014 y 2016. Además, los porcentajes que se aproximan al 35% que se indican en las variables, corresponden a que los estudiantes sólo aplican dos de las cuatro competencias genéricas mencionadas en el capítulo II.

De acuerdo a lo anterior se procedió a la eliminación de variables como se muestra en la tabla 31.

Tabla 31 Consolidado de datos anómalos o nulos

Variable	Descripción	Cantidad de datos eliminados
Jornada	Para algunos estudiantes aparecía la jornada como un código numérico al cual no se le encontró una sede relacionada. Se eliminaron aquellos datos que no especificaban la en la tabla de sedes.	79630
Establecimientos	Para algunos estudiantes aparecía el establecimiento como un código alfanumérico al cual no se le encontró una sede relacionada. Se eliminaron los registros que no fueron encontrados en las bases 2015 y 2016.	79630
Leng_score	Hace referencia a la competencia de lenguaje y está dividida en cinco variables (leng_score1, leng_score2, leng_score3,	57383

	leng_score4, leng_score5) y sus valores se sobrepasan del rango especificado (-3 a 3)[3] en el informe técnico dado por el ICFES, (ICFES, 2011). Además, poseían celdas con datos nulos o vacíos.	
Mate_score	Se refiere a la competencia de lenguaje y está dividida en cinco variables (mate_score1, mate _score2, mate _score3, mate _score4, mate _score5) y sus valores se sobrepasan del rango especificado (-3 a 3) en el informe técnico dado por el ICFES, (ICFES, 2011). Además, poseían celdas con datos nulos o vacíos.	81768
Cien_score	Hace referencia a la competencia de lenguaje y está dividida en cinco variables (cien_score1, cien _score2, cien _score3, cien _score4, cien _score5) y sus valores se sobrepasan del rango especificado (-3 a 3) en el informe técnico dado por el ICFES, (ICFES, 2011). Además, poseían celdas con	63435

[3] "Para los estudiantes que presentaron la prueba SABER 3°, 5°. y 9°, se obtuvo el resultado de su desempeño a partir de cinco valores, denominados valores plausibles (PV1 a PV5) o como están en el FTP (score1 a score5) que generalmente vienen en una escala de -3 a 3. Estos valores son útiles, pues tienen en cuenta la aleatoriedad producida por el hecho de que los estudiantes responden a un número pequeño de preguntas, lo cual permite obtener mejores estimaciones de las estadísticas de interés relacionadas con el desempeño en las pruebas a nivel agregado".(ICFES, 2011)

	datos nulos o vacíos.	
Comp_score	Hace referencia a la competencia de lenguaje y está dividida en cinco variables (cien_score1, cien _score2, cien _score3, cien _score4, cien _score5) y sus valores se sobrepasan del rango especificado (-3 a 3) en el informe técnico dado por el ICFES, (ICFES, 2011). Además, poseían celdas con datos nulos o vacíos.	29406
Aplicación	Contenía valores constantes ya que todos los estudiantes de las tres bases de datos aplicaron la prueba.	2082451
Estu_consecutivo	Contenía valores constantes puesto que solo sirve como identificadores de los estudiantes.	2082451
Grupo	Contenía valores constantes puesto que es un identificador de las pruebas censal y control. En este caso todos los estudiantes presentaron la prueba censal.	2082451
Estrato	Contenía valores alfanuméricos. Existen otras variables en las cuales está contenida como son establecimiento, nivel socioeconómico, sector, zona y grupo.	2082451

N	Solo contenía el número de participantes que presentaron la prueba por establecimiento o por sede.	2082451
Grado	Es un valor constante puesto que la prueba es de todos los estudiantes de grado 5°.	2082451
Modelo educativo	Los establecimientos no tienen definido un modelo educativo en general.	2082451
Dissenso	Contiene datos de estudiantes que presentan o no alguna discapacidad. Se elimina ya que el número de estudiantes que presentan discapacidad es muy pequeño y puede generar sesgo en el análisis.	2082451

Nota. Elaboración propia

Para darle solución al porcentaje de valores nulos indicados en la tabla anterior se dividió los datos en tablas según las competencias que se hayan aplicado en determinado año en las Pruebas Saber.

La tabla 32 contiene 16 filas cada una de ellas corresponde a las posibles combinaciones de las competencias genéricas que se pueden dar en su aplicación. Se obtuvieron un total de 16 tablas, de las cuales se omitieron siete debido a que su combinación no se podía dar.

Tabla 32 Consolidado de las posibles combinaciones entre competencias genéricas

N° De Fila	Lenguaje	Mate	Ciencias	Comp	Número De Datos
1	0	0	0	0	0
2	0	0	0	1	2170
3	0	0	1	0	4733
4	0	0	1	1	0
5	0	1	0	0	10075
6	0	1	0	1	223876
7	0	1	1	0	458142
8	0	1	1	1	0
9	1	0	0	0	11121
10	1	0	0	1	223466
11	1	0	1	0	459708
12	1	0	1	1	0
13	1	1	0	0	689160
14	1	1	0	1	0
15	1	1	1	0	0
16	1	1	1	1	0

Nota. Elaboración propia

La fila 1 corresponde a la tabla que muestra que no se aplicó ninguna competencia, lo que no es posible, porque en los años estudiados deben aplicarse tres de las cuatro competencias genéricas.

Las filas 2, 3, 5 y 9 se relacionan con las tablas que aparecen con una sola competencia aplicada y corresponden a competencias ciudadanas (2170 datos), Ciencias naturales (4733 datos), matemáticas (10075 datos), lenguaje (11121 datos) respectivamente. Estos datos se obviaron en el proceso puesto que no cumple con las condiciones iniciales planteadas en los lineamientos de las pruebas.

La fila 4 se obvio puesto que la combinación de las competencias ciudadanas y de ciencias naturales no se aplican simultáneamente si no en años separados y sumado a esto no contiene datos.

La fila 8 muestra la aplicación de las competencias: matemáticas, ciudadanas y ciencias naturales, lo cual no es posible por lo mencionado en el párrafo anterior y por esto no contiene valores.

La fila 12 muestra la aplicación de tres competencias: lenguaje, ciudadanas y ciencias naturales, lo cual no es posible por lo mencionado anteriormente y por esto no contiene valores.

La fila 14 y 15 se obvio porque presenta la combinación de tres competencias y esto no se puede dar.

La fila 16 se excluye puesto que no se puede aplicar las cuatro competencias al mismo tiempo.

Finalmente, solo quedaron las filas 6, 7, 10,11 y 13 que corresponden a la aplicación de competencias ciudadanas y matemáticas, ciencias naturales y matemáticas, competencias ciudadanas y lenguaje, ciencias naturales y lenguaje, y matemáticas y lenguaje, respectivamente; Con cada una de estas filas se construyó una tabla que conservan los atributos de la tabla de estudiantes.

3.3.2 Transformación

Se sabe que la alta dimensionalidad es un problema para el descubrimiento de patrones con minería de datos (Hernández et al., 2005). Uno de los criterios utilizados para resolver este problema es la reducción del número de atributos por analizar, a través de su transformación en nuevos atributos que generalicen los datos y que ofrezcan mayor información. Teniendo en cuenta este criterio, en el repositorio de datos saber5_2014_2016 se omitieron aquellos atributos que por sí mismos presentaban inconsistencias como por ejemplo, los SCORE, que se encontraban desfasados respecto al rango establecido en el informe técnico dado por el ICFES, (2011). Los SCORE son estimaciones para cada competencia que constan de 5 valores (PV1 a PV5), los cuales se promedian para determinar la estimación del puntaje promedio, dicho puntaje promedio es el atributo WEIGHT, con el cual se trabajará en el nuevo repositorio de datos.

La variable WEIGHT toma un rango de valores diferente en cada año como se puede observar en la tabla 33, por lo cual se decidió redondear los datos a valores enteros y normalizarlos en una escala de 1 a 4 con el fin de hacer la discretización según los niveles de desempeño (insuficiente, mínimo, satisfactorio, avanzado), los cuales se encuentran especificados en la tabla 35. Además, se integraron algunos atributos de la tabla de los establecimientos al repositorio Saber5_2014_2016 para obtener mayor información sobre las características de los estudiantes.

Tabla 33 Resumen Estadístico para WEIGHT

Año	Recuento	Promedio	Desviación Estándar	Coeficiente de Variación	Mínimo	Máximo	Rango
2014	486090	4,36467	20,6405	472,90%	1	400	399,0
2015	453557	1,71537	0,269975	15,74%	1	6	5
2016	443808	1,59989	0,210674	13,17%	1	6	5
Total	1383455	2,60918	2,3044	471,58%	1	400	399

Nota. Elaboración propia

En la tabla 34 se describe el proceso de transformación de las variables contenidas en el nuevo repositorio de datos, donde se muestran aquellos atributos que se incluyen y/o se reemplazan.

Tabla 34 Nuevo diccionario de datos del repositorio saber5_2014_2016

Nombre De Campo	Descripción	Acción	Valores
Jornada	(Codsitio) - Código de la sede-jornada al que pertenece el estudiantes	Se cambia el valor numérico a carácter	C:Completa M: Mañana T: Tarde U: Única
Establecimiento	Código DANE del establecimiento educativo al que pertenece el estudiante	se reemplaza el código por el nombre de la institución	Llave a tabla INSTITUCION
EnteTerr	Código DANE de la entidad a la que pertenece establecimiento educativo	Se cambia el código por el nombre de la entidad territorial	Llave a tabla ENTIDADTERRITORIAL
Departamento	Código DANE del departamento al que pertenece	Se cambia el código del departamento	Llave a tabla DEPARTAMENTO

	establecimiento educativo	por el nombre	
Municipio	Código DANE del municipio al que pertenece establecimiento educativo	Se cambia el código del municipio por el nombre	Llave a tabla MUNICIPIO
Género	Sexo del estudiante	Se reemplaza el sexo por género	Masculino Femenino No especifica
Sector	Sector del establecimiento	Se cambia el valor numérico por carácter	Oficial No oficial
Calendario	Calendario del establecimiento	No se realizó ninguna acción	A , B
Nivel Socioeconómico	Nivel socioeconómico del establecimiento	No se realizó ninguna acción	1 , 2 , 3 , 4 , 5
Leng_copietas	Indicador de copia de lenguaje	Se cambia el valor numérico por carácter	no copia con copia
Leng_weight	Peso muestral de lenguaje	Se reemplaza a los atributos leng_Score1,	Insuficiente mínimo satisfactorio

		leng_Score2, leng_Score3, leng_Score4, leng_Score5	avanzado
Mate_copietas	Indicador de copia de matemáticas	Se cambia el valor numérico por carácter	no copia con copia
Mate_weight	Peso muestral de matemáticas	Se reemplaza a los atributos mate_Score1, mate_Score2, mate_Score3, mate_Score4, mate_Score5	insuficiente mínimo satisfactorio avanzado
Cien_copietas	Indicador de copia de ciencias	Se cambia el valor numérico por carácter	no copia con copia
Cien_weight	Peso muestral de ciencias	Se reemplaza a los atributos cien_Score1, cien _Score2, cien _Score3, cien _Score4,	insuficiente mínimo satisfactorio avanzado

		cien _Score5	
Comp_copietas	Indicador de copia de competencias ciudadanas	Se cambia el valor numérico por carácter	no copia con copia
Comp_weight	Peso muestral de competencias ciudadanas	Se reemplaza a los atributos comp_Score1, comp _Score2, comp _Score3, comp _Score4, comp _Score5	Insuficiente Mínimo Satisfactorio Avanzado
Zona	Zona donde se ubica la mayoría de la población atendida por el establecimiento educativo	Se cambia el valor numérico (1, 2) por carácter	Urbano Rural
Sector	Naturaleza administrativa del establecimiento educativo	Se cambia el valor numérico (1,2) por carácter	Oficial No oficial
Tipo_estab	Tipo de establecimiento	Se cambia el valor numérico (1, 2, 3) por	Oficial urbano Oficial rural No Oficial

		carácter	

Nota. Elaboración propia

Para facilitar la búsqueda de patrones de rendimiento académico, se discretizaron los valores numéricos de ciertos atributos, para esto se tuvo en cuenta un rango de valores y la proporcionalidad de las frecuencias por cada valor con el fin de evitar sesgos en la construcción de modelos de la Minería de Datos. En la tabla 35 se evidencia el proceso de normalización para el atributo *Weight* para cada una de las competencias genéricas, donde se tiene en cuenta los niveles de desempeño dados por el ICFES.

Tabla 35 Discretización de valores del atributo Weight.

Año	**Competencias Genéricas**	**Descripción inicial**	**Normalización**	**Estado actual**
2014	Lenguaje Matemáticas Ciencias naturales Competencias ciudadanas	Toma valores decimales en un rango de 1 hasta 400.	Se aproxima los números decimales a enteros y se discretiza los valores en un rango de 1 a 4.	1= Insuficiente 2= Mínimo 3= Satisfactorio 4= Avanzado
2015	Lenguaje Matemáticas Ciencias naturales	Toma valores decimales en un rango de 1 hasta 6.	Se aproxima los números decimales a enteros y se	1= Insuficiente 2= Mínimo 3= Satisfactorio 4= Avanzado

	Competencias ciudadanas		discretiza los valores en un rango de 1 a 4.	
2016	Lenguaje Matemáticas Ciencias naturales Competencias ciudadanas	Toma valores decimales en un rango de 1 hasta 6.	Se aproxima los números decimales a enteros y se discretiza los valores en un rango de 1 a 4.	1= Insuficiente 2= Mínimo 3= Satisfactorio 4= Avanzado

Nota. Elaboración propia

Para hacer un estudio en profundidad con respecto al rendimiento académico en competencias genéricas se consideró pertinente crear un nuevo atributo denominado zonas en el cual se reorganizan las instituciones educativas según zonas geográficas, de manera que la información suministrada aporte al descubrimiento de desempeño académico.

Durante el proceso se consideraron varias opciones para la clasificación de dichas zonas, entre las cuales están la reorganización por zonas geográficas como tal, zonas político-administrativas, zonas culturales y zonas de los Órganos Colegiados de Administración y Decisión (OCAD). Los OCAD se constituyen a nivel municipal, departamental, regional, nacional.

Estas zonas OCAD, en la actualidad, constituyen un reordenamiento territorial del país frente a las regiones geográficas que ha manejado históricamente y que ha permitido marcar diferencias significativas entre el centro y la periferia de Colombia. La ventaja de esta organización es que cada zona OCAD agrupa departamentos completos, con sus respectivos municipios. Gracias

a esto es posible predecir cuál sería el rendimiento académico en las competencias genéricas en estas zonas con base en la información de las pruebas Saber 5.

Los OCAD se encuentran distribuidos en seis zonas que cubren la totalidad del país. Teniendo en cuenta que Bogotá es la ciudad donde se concentran la gran mayoría de instituciones, para esta investigación se consideró conveniente tomarla como una nueva zona. Así pues, para este estudio las zonas geográficas en las que se agrupan las instituciones educativas son siete, como se detallan en la tabla 36 (Timarán, S., et al, 2016).

Tabla 36 Valores del atributo zona

Zonas	Departamentos
Caribe	Atlántico, Bolívar, Cesar, Córdoba, Guajira, Magdalena, San Andrés, Providencia y Santa Catalina y Sucre.
Centro sur	Amazonas, Caquetá, Huila, Putumayo y Tolima.
Centro oriente	Boyacá, Cundinamarca, Norte de Santander y Santander.
Eje cafetero	Antioquia, Caldas, Risaralda y Quindío.
Llano	Arauca, Casanare, Guainía, Guaviare, Meta, Vaupés y Vichada.
Pacífico	Cauca, Chocó, Nariño y Valle del Cauca.
Bogotá	Distrito Capital de Bogotá

Fuente. Adaptado de Colombia. Departamento Nacional de Planeación. Sistema General de Regalías sgr. (2012). Bogotá: DNP.

Clasificadas las instituciones educativas por zonas se procedió a contar el número de estudiantes e instituciones educativas por zona con lo cual se crearon dos nuevos atributos denominados num_estu_zona y num_inst_zona.

Para los valores del atributo num_estu_zona, se construyen tres clases y se cuenta la frecuencia de cada intervalo y luego se discretizan en tres categorías donde se considera alto si el

número de estudiantes es mayor de 300.000, medio si está entre 200.000 y 300.000 y bajo si es menor que 200.000. En la tabla 37 se muestran los valores de referencia.

Tabla 37 Valores discretizados del atributo num_estu_zona

Zona	N° estudiantes	Valor
Caribe	506198	Alto
Centro sur	160236	Bajo
Centro oriente	335555	Medio
Eje cafetero	344754	Medio
Llano	93936	Bajo
Pacífico	314808	Medio
Bogotá	282633	Medio
Total de estudiantes	2038120	

Nota. Elaboración propia

Para los valores del atributo num_inst_zona, se considera alto si el número de instituciones educativas es mayor que 4000, medio si el número de instituciones educativas está entre 2000 y 4000, y bajo si es menor que 2000. En la tabla 38 se muestran los resultados de la discretización.

Tabla 38 Valores discretizados del atributo num_inst_zona

Zona	**N° instituciones**	**valor**
Caribe	4863	Alto
Centro sur	1594	Bajo
Centro oriente	3204	Medio
Eje cafetero	5434	Alto
Llano	865	Bajo

Pacífico	5761	Alto
Bogotá	1998	Bajo
Total instituciones	23719	

Nota. Elaboración propia

Como resultado de todos los procesos descritos, se obtuvo un nuevo repositorio de datos limpio y transformado, listo para aplicarle las técnicas de minería de datos. En la tabla 39 se describen el diccionario de datos del repositorio minable, el cual contiene 2’038.120 registros y 18 atributos, el cual se encuentra organizado por los factores socioeconómicos, académicos e institucionales.

Tabla 39 Diccionario de datos del repositorio final

N°	Atributo	Descripción	Valores
SOCIOECONÓMICOS			
1	Género	Sexo del estudiante	Femenino Masculino No especifica
2	Nivel socioeconómico	Nivel socioeconómico del establecimiento educativo	1, 2, 3, 4 y 5
3	Departamento	Departamento al que pertenece el establecimiento educativo	Nombre del departamento
4	Municipio	Municipio al que pertenece el establecimiento educativo	Nombre del municipio
5	Zona	Zona donde se ubica la mayoría de la población atendida	Urbana Rural

ACADÉMICAS			
✓ **6**	Lenguaje	Desempeño académico del estudiante en la competencia genérica lenguaje	Insuficiente Mínimo Satisfactorio Avanzado
7	Matemáticas	Desempeño académico del estudiante en la competencia genérica matemáticas	Insuficiente Mínimo Satisfactorio Avanzado
8	Ciencias Naturales	Desempeño académico del estudiante en la competencia genérica ciencias	Insuficiente Mínimo Satisfactorio Avanzado
9	Competencias Ciudadanas	Desempeño académico del estudiante en la competencia genérica competencias	Insuficiente Mínimo Satisfactorio Avanzado
10	Leng_copietas	Indicador de copia de lenguaje	Copia y No copia
11	Mate_copietas	Indicador de copia de matemáticas	Copia y No copia
12	Cien_copietas	Indicador de copia de ciencias	Copia y No copia
13	Comp_copietas	Indicador de copia de competencias	Copia y No copia
INSTITUCIONALES			

14	Sector	Sector del establecimiento educativo	Oficial y No oficial
15	Tipoent	Tipo de entidad territorial a la que pertenece el establecimiento educativo	No certificadas y Certificadas
16	Calendario	Calendario del establecimiento educativo	A y B
17	num_estu_zona	Número de estudiantes por zonas geográficas	Alto, Medio y Bajo
18	num_inst_zona	Número de instituciones por zonas geográficas	Alto, Medio y bajo
19	Jornada	Jornada a la que pertenece el estudiante	M, T, C y U
20	Zonaestab	Zona del establecimiento educativo	Urbano y Rural

Nota. Elaboración propia

Según los parámetros del ICFES la prueba evalúa dos de las cuatro competencias genéricas a cada estudiante, se construyeron cinco tablas, una por cada par de competencias genéricas a partir del repositorio de datos saber5_2014_2016. En la tabla 40 se muestran las características de este conjunto de datos que se constituyen en los repositorios minables para aplicar las técnicas descriptivas de Minería de Datos.

Tabla 40 Conjuntos de datos por competencias genéricas

Repositorio	N° registros	N° atributos	Descripción

minable_mate_comp	222518	18	Conjunto de datos de estudiantes que presentaron prueba de matemáticas y competencias ciudadanas
minable_mate_cien	458121	18	Conjunto de datos de estudiantes que presentaron prueba de matemáticas y ciencias naturales
minable_leng_comp	229748	18	Conjunto de datos de estudiantes que presentaron prueba de lenguaje y competencias ciudadanas
minable_leng_cien	449823	18	Conjunto de datos de estudiantes que presentaron prueba de lenguaje y ciencias naturales
minable_leng_mate	677910	18	Conjunto de datos de estudiantes que presentaron prueba de lenguaje y matemáticas

Nota. Elaboración propia

3.4 Modelado

Teniendo en cuenta que el objetivo de esta investigación es detectar patrones que inciden en el rendimiento de las Pruebas saber 5, en los años 2014 a 2016 a nivel nacional en los aspectos socioeconómicos, académicos e institucionales con el fin de generar conocimiento encaminado a soportar las decisiones institucionales y gubernamentales para el mejoramiento de la calidad educativa, se exploraron y evaluaron las tareas descriptivas de minería de datos. En esta fase se seleccionó la tarea de reglas de asociación con el algoritmo Apriori (Agrawal, Imielinski & Swami, 1993) y de agrupamiento o clustering con el algoritmo k-means, ya que estos algoritmos se ajustan al propósito del objetico de esta investigacion. Con la asociación se pretende identificar cuáles son los factores que están sucediendo frecuentemente juntos y que están relacionados con el rendimiento académico en las pruebas mencionadas anteriormente y con agrupación se busca obtener grupos homogéneos de estudiantes que presentan similitudes en los factores asociados al mejor rendimiento.

Se escogió la herramienta Weka (Waikato Environment for Knowledge Analysis) (García, s. f.) fue desarrollada en la Universidad de Waikato (Nueva Zelanda) bajo licencia GPL (General Public License). Esta herramienta permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario. Weka es una de las suites más utilizadas en el área de descubrimiento de conocimiento en los últimos años.

3.4.1 Tarea de Asociación

En la base de datos saber5 2014_2016 se cuenta con un conjunto de atributos y una colección de registros, con los cuales se busca encontrar relaciones para descubrir reglas de asociación que cumplan con unas especificaciones mínimas expresadas en forma de soporte y

confianza. Para esto se calculó el promedio de las dos variables dependientes, correspondientes a las competencias genéricas, que hacen parte de cada repositorio minable (ver tabla 32), con el objetivo de obtener una sola variable dependiente y pueda ejecutarse el algoritmo. Tomando lo anterior se extrajeron reglas que determinaron ciertas características que aparecen juntas según el nivel de desempeño.

La herramienta de minería de datos weka que contiene la implementación del algoritmo de aprendizaje de reglas de asociación Apriori se pude configurar con varias opciones (ver figura 7): con la opción UpperBoundMinSupport se indica el límite inferior de soporte requerido para aceptar un conjunto de ítems. Si no se encuentran conjuntos de ítems suficientes para generar las reglas requeridas se va disminuyendo el límite hasta llegar al límite inferior (opción LowerBoundMinSupport). Con la opción minMetric se indica la confianza mínima (u otras métricas dependiendo del criterio de ordenación) para mostrar una regla de asociación; y con la opción numRules se indica el número de reglas que se desea generar. El ordenamiento de estas reglas se puede configurar mediante la opción MetricType, algunas opciones que se pueden utilizar son: confianza de la regla, lift (confianza divido por el número de ejemplos cubiertos por la parte derecha de la regla), y otras más elaboradas (Hernández & Ferri, 2006).

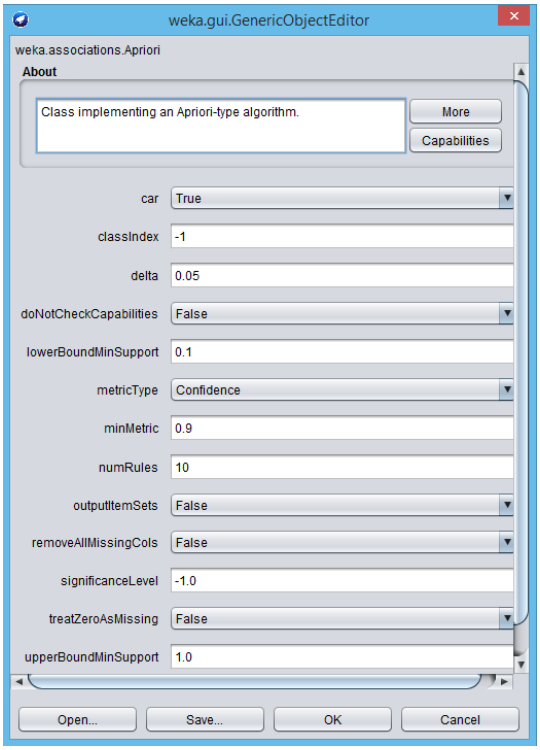

Figura 7 Vista de configuración del algoritmo Apriori en Weka

Para generar las reglas de asociación se creyó conveniente aplicar el algoritmo Apriori con el fin de obtener las más fuertes, ya que las reglas de asociación permiten descubrir factores asociados al desempeño en las competencias genéricas de las Pruebas Saber 5.

En la tabla 41 se especifican los atributos comunes a los repositorios de datos de la tabla 40, estas variables son las que se usan para le ejecución de los algoritmos que se describe más adelante.

Tabla 41 Atributos comunes a los repositorios minables

Atributo	Descripción
Género	Sexo del estudiante
Nivel socioeconómico	Nivel socioeconómico del establecimiento educativo.
Sector	Sector del establecimiento educativo

Zona	Zona donde se ubica la mayoría de la población atendida
Zonaestab	Zona del establecimiento educativo
Tipoent	Tipo de entidad territorial a la que pertenece el establecimiento educativo
Sectorestab	Naturaleza administrativa del establecimiento educativo
Tipoestab	Tipo de establecimiento
Región	Región a la que pertenece el estudiante
Num_estu_región	Número de estudiantes por región
Num_inst_región	Número de instituciones por región
Leng_copietas	Indicador de copia de lenguaje
Comp_copietas	Indicador de copia de competencias ciudadanas
Mate_copietas	Indicador de copias de matemáticas
Cien_copietas	Indicador de copias de ciencias naturales
P_leng	Desempeño académico del estudiante en la competencia genérica de matemáticas
P_comp	Desempeño académico del estudiante en la competencia genérica de competencias ciudadanas
P_cien	Desempeño académico del estudiante en la competencia genérica de ciencias naturales
P_mate	Desempeño académico del estudiante en la competencia genérica de matemáticas

Fuente. Tomado de las bases de datos ICFES.

3.4.1.1 Reglas Generadas con el Algoritmo A priori

- ✓ **Competencias Genéricas de Lenguaje y Ciencias Naturales (Lengcien)**

Para obtener las reglas fuertes hubo la necesidad de seleccionar ciertos atributos para el experimento y omitir otros que presentaban las mismas características. Se estableció diferentes valores para los parámetros del soporte y la confianza pero los que generaron las mejores reglas tienen como mínima confianza el 92% (0.92), un soporte mínimo superior de 1.0, un soporte mínimo inferior de 0.2, un Δ de 0.5 y un número de reglas a generar de 10.000. También se filtraron las reglas para obtener solo aquellas, donde el atributo lengcien se encuentre como consecuente de la regla. Las reglas fuertes resultantes fueron aquellas con un soporte mínimo del 20% (0.2). Los parámetros de ejecución del algoritmo Apriori se presentan en la figura 8 y las mejores 24 reglas generadas con una confianza del 92%,

```
=== Run information ===
Scheme:       weka.associations.Apriori -N 100 -T 0 -C 0.92 -D 0.05 -U
1.0 -M 0.2 -S -1.0 -A -c 5
Relation:     minable_leng_cienl-
weka.filters.unsupervised.attribute.Remove-R3-4,7-11,15,20
Instances:    445699
Attributes:   11
              jornada
              genero
              zonaestab
              nivelsocioeconomico
              lengcien
              tipoent
              sectoresta
              calendario
              region
              num_estu_zona
              num_inst_zona
=== Associator model (full training set) ===

Apriori
=======

Minimum support: 0.2 (89140 instances)
Minimum metric <confidence>: 0.92
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16
Size of set of large itemsets L(2): 64
Size of set of large itemsets L(3): 93
Size of set of large itemsets L(4): 47
Size of set of large itemsets L(5): 3
```

Figura 8 Parámetros de ejecución del algoritmo Apriori para el repositorio Lenguaje y Ciencias Naturales.

Best rules found:

1. jornada=T sectoresta=Oficial 99260 ==> lengcien=Minimo 93318 conf:(0.94)
2. jornada=T sectoresta=Oficial calendario=A 99260 ==> lengcien=Minimo 93318 conf:(0.94)
3. zonaestab=Urbano tipoent=no certificadas sectoresta=Oficial calendario=A 145604 ==> lengcien=Minimo 136718 conf:(0.94)
4. zonaestab=Urbano sectoresta=Oficial calendario=A num_inst_zona=alto 146769 ==> lengcien=Minimo 137489 conf:(0.94)
5. zonaestab=Urbano sectoresta=Oficial num_inst_zona=alto 146779 ==> lengcien=Minimo 137497 conf:(0.94)
6. zonaestab=Urbano nivelsocioeconomico=medio sectoresta=Oficial 114159 ==> lengcien=Minimo 106927 conf:(0.94)
7. zonaestab=Urbano tipoent=no certificadas sectoresta=Oficial calendario=A num_estu_zona=medio 111447 ==> lengcien=Minimo 104376 conf:(0.94)
8. zonaestab=Urbano tipoent=no certificadas sectoresta=Oficial num_estu_zona=medio 111457 ==> lengcien=Minimo 104384 conf:(0.94)
9. nivelsocioeconomico=medio sectoresta=Oficial 117034 ==> lengcien=Minimo 109508 conf:(0.94)
10. jornada=T 107351 ==> lengcien=Minimo 100296 conf:(0.93)
11. jornada=T calendario=A 107054 ==> lengcien=Minimo 100012 conf:(0.93)
12. jornada=T zonaestab=Urbano 97989 ==> lengcien=Minimo 91538 conf:(0.93)
13. jornada=T zonaestab=Urbano calendario=A 97695 ==> lengcien=Minimo 91257 conf:(0.93)
14. tipoent=no certificadas sectoresta=Oficial 166044 ==> lengcien=Minimo 154387 conf:(0.93)
15. tipoent=no certificadas sectoresta=Oficial num_estu_zona=medio 122321 ==> lengcien=Minimo 113717 conf:(0.93)
16. zonaestab=Urbano nivelsocioeconomico=bajo calendario=A num_inst_zona=alto 100761 ==> lengcien=Minimo 93397 conf:(0.93)
17. zonaestab=Urbano nivelsocioeconomico=bajo num_inst_zona=alto 101135 ==> lengcien=Minimo 93693 conf:(0.93)
18. genero=Femenino zonaestab=Urbano sectoresta=Oficial calendario=A 128053 ==> lengcien=Minimo 118607 conf:(0.93)
19. zonaestab=Urbano sectoresta=Oficial calendario=A num_estu_zona=medio 168430 ==> lengcien=Minimo 155827 conf:(0.93)
20. zonaestab=Urbano sectoresta=Oficial num_estu_zona=medio 168440 ==> lengcien=Minimo 155835 conf:(0.93)
21. genero=Masculino zonaestab=Urbano sectoresta=Oficial calendario=A 126376 ==> lengcien=Minimo 116875 conf:(0.92)
22. genero=Masculino zonaestab=Urbano sectoresta=Oficial 126380 ==> lengcien=Minimo 116878 conf:(0.92)
23. zonaestab=Urbano calendario=A num_estu_zona=medio num_inst_zona=alto 105091 ==> lengcien=Minimo 96992 conf:(0.92)
24. jornada=M zonaestab=Urbano sectoresta=Oficial calendario=A 153160 ==> lengcien=Minimo 141280 conf:(0.92)

Figura 9 Mejores reglas generadas con Apriori con el conjunto de datos de Lenguaje y Ciencias Naturales.

✓ **Competencias Genéricas de Lenguaje y Competencias Ciudadanas (Lengcomp)**

En el caso del repositorio de datos lengcomp. Se establecieron diferentes valores para los parámetros del soporte y la confianza pero los que generaron las reglas fuertes tienen como mínima confianza el 96% (0.96), un soporte mínimo superior de 1.0, un soporte mínimo inferior de 0.25, un Δ de 0.5 y un número de reglas a generar de 10.000. También se filtraron las reglas para obtener solo aquellas, donde el atributo lengcomp se encuentre como consecuente de la regla. Las reglas fuertes fueron aquellas con un soporte mínimo del 25% (0.25). Los parámetros de ejecución del algoritmo Apriori se presentan en la figura 10 y las mejores 24 reglas generadas con una confianza del 96%, en la figura 11.

```
=== Run information ===
Scheme:weka.associations.Apriori -N 100 -T 0 -C 0.96 -D 0.05 -U 1.0 -M 0.25
-S -1.0 -A -c 5
Relation:     minable_leng_compl-weka.filters.unsupervised.attribute.Remove-R3-4,7-
11,20-weka.filters.unsupervised.attribute.Remove-R8
Instances:    219748
Attributes:   11
              jornada
              genero
              zonastab
              nivelsocioeconomico
              lengcomp
              tipoent
              sectoresta
              calendario
              region
              num_estu_zona
              num_inst_zona
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.25 (54937 instances)
Minimum metric <confidence>: 0.96
Number of cycles performed: 15
Generated sets of large itemsets:
Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 49
Size of set of large itemsets L(3): 63
Size of set of large itemsets L(4): 25
Size of set of large itemsets L(5): 2
```

Figura 10 . Parámetros de ejecución del algoritmo Apriori para el repositorio Lenguaje y competencias Ciudadanas.

```
 Best rules found:
1. zonastab=Urbano tipoent=no certificadas sectoresta=Oficial 70736 ==>
   lengcomp=Minimo 68558    conf:(0.97)
2. zonastab=Urbano tipoent=no certificadas sectoresta=Oficial calendario=A 70720 ==>
   lengcomp=Minimo 68542    conf:(0.97)
3. zonastab=Urbano nivelsocioeconomico=medio sectoresta=Oficial calendario=A
   num_estu_zona=medio 57130 ==> lengcomp=Minimo 55327    conf:(0.97)
4. nivelsocioeconomico=medio sectoresta=Oficial num_estu_zona=medio 58833 ==>
   lengcomp=Minimo 56946    conf:(0.97)
5. nivelsocioeconomico=medio sectoresta=Oficial calendario=A num_estu_zona=medio
   58833 ==> lengcomp=Minimo 56946    conf:(0.97)
6. zonastab=Urbano nivelsocioeconomico=medio sectoresta=Oficial calendario=A 74698
   ==> lengcomp=Minimo 72247    conf:(0.97)
7. nivelsocioeconomico=medio sectoresta=Oficial calendario=A 76643 ==>
   lengcomp=Minimo 74101    conf:(0.97)
8. nivelsocioeconomico=medio tipoent=no certificadas calendario=A
   num_estu_zona=medio 66503 ==> lengcomp=Minimo 64226    conf:(0.97)
9. zonastab=Urbano nivelsocioeconomico=medio tipoent=no certificadas calendario=A
   num_estu_zona=medio 65243 ==> lengcomp=Minimo 63006    conf:(0.97)
10. zonastab=Urbano nivelsocioeconomico=medio tipoent=no certificadas
   num_estu_zona=medio 68236 ==> lengcomp=Minimo 65840    conf:(0.96)
11. nivelsocioeconomico=medio tipoent=no certificadas calendario=A 79991 ==>
   lengcomp=Minimo 77153    conf:(0.96)
12. zonastab=Urbano nivelsocioeconomico=medio tipoent=no certificadas calendario=A
   78625 ==> lengcomp=Minimo 75828    conf:(0.96)
13. nivelsocioeconomico=medio tipoent=no certificadas 83290 ==> lengcomp=Minimo 80274
   conf:(0.96)
14. zonastab=Urbano nivelsocioeconomico=medio calendario=A num_estu_zona=medio 80819
   ==> lengcomp=Minimo 77889    conf:(0.96)
15. zonastab=Urbano nivelsocioeconomico=medio num_estu_zona=medio 83932 ==>
   lengcomp=Minimo 80839    conf:(0.96)
16. zonastab=Urbano tipoent=no certificadas calendario=A 100835 ==> lengcomp=Minimo
   97099    conf:(0.96)
17. zonastab=Urbano tipoent=no certificadas calendario=A num_estu_zona=medio 77146
   ==> lengcomp=Minimo 74262    conf:(0.96)
18. zonastab=Urbano tipoent=no certificadas num_estu_zona=medio 80338 ==>
   lengcomp=Minimo 77290    conf:(0.96)
19. genero=Femenino zonastab=Urbano sectoresta=Oficial calendario=A 63088 ==>
   lengcomp=Minimo 60626    conf:(0.96)
20. tipoent=no certificadas sectoresta=Oficial calendario=A num_estu_zona=medio 59461
   ==> lengcomp=Minimo 57130    conf:(0.96)
21. tipoent=no certificadas sectoresta=Oficial num_estu_zona=medio 59503 ==>
   lengcomp=Minimo 57155    conf:(0.96)
22. jornada=M zonastab=Urbano sectoresta=Oficial 81709 ==> lengcomp=Minimo 78468
   conf:(0.96)
23. jornada=M zonastab=Urbano sectoresta=Oficial calendario=A 81693 ==>
   lengcomp=Minimo 78452    conf:(0.96)
24. zonastab=Urbano nivelsocioeconomico=medio calendario=A 106821 ==> lengcomp=Minimo
   102554    conf:(0.96)
```

Figura 11 Mejores reglas generadas con Apriori con el conjunto de datos de Lenguaje y Competencias Ciudadanas.

✓ **Competencias Genéricas de Lenguaje y Matemáticas (Lengmate)**

De igual manera, se obtuvieron reglas de asociación que relacionan solo los factores para la prueba de Lenguaje y Matemáticas. Se establecieron diferentes valores para los parámetros del soporte y la confianza pero los que generaron las reglas fuertes tienen como mínima confianza el 91% (0.91), un soporte mínimo superior de 1.0, un soporte mínimo inferior de 0.25, un Δ de 0.5 y un número de reglas a generar de 10.000. También se filtraron las reglas para obtener solo aquellas, donde el atributo lengmate se encuentre como consecuente de la regla. Las reglas fuertes fueron aquellas con un soporte mínimo del 25% (0.25). Los parámetros de ejecución del algoritmo Apriori se presentan en la figura 12 y las mejores 24 reglas generadas con una confianza del 91%, en la figura 13.

```
=== Run information ===
Scheme: weka.associations.Apriori -N 100 -T 0 -C 0.91 -D 0.05 -U 1.0
-M 0.25 -S -1.0 -A -c 5
Relation:     minable_leng_matel-weka.filters.unsupervised.attribute.Remove-R3-4,7-
11,15,20
Instances:    671052
Attributes:   11
              jornada
              genero
              zonastab
              nivelsocioeconomico
              lengmate
              tipoent
              sectoresta
              calendario
              region
              num_estu_zona
              num_inst_zona
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.25 (167763 instances)
Minimum metric <confidence>: 0.91
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 45
Size of set of large itemsets L(3): 52
Size of set of large itemsets L(4): 20
```

Figura 12 Parámetros de ejecución del algoritmo Apriori para el repositorio Lenguaje y Matemáticas.

```
Best rules found:

   1. zonastab=Urbano tipoent=no certificadas sectoresta=Oficial calendario=A
      217327 ==> lengmate=Minimo 204172    conf:(0.94)
   2. zonastab=Urbano nivelsocioeconomico=medio sectoresta=Oficial 189545 ==>
      lengmate=Minimo 176774    conf:(0.93)
   3. zonastab=Urbano nivelsocioeconomico=medio sectoresta=Oficial calendario=A
      189545 ==> lengmate=Minimo 176774    conf:(0.93)
   4. nivelsocioeconomico=medio sectoresta=Oficial calendario=A 194342 ==>
      lengmate=Minimo 181028    conf:(0.93)
   5. tipoent=no certificadas sectoresta=Oficial calendario=A num_estu_zona=medio
      182471 ==> lengmate=Minimo 169872    conf:(0.93)
   6. tipoent=no certificadas sectoresta=Oficial 248229 ==> lengmate=Minimo 231006
      conf:(0.93)
   7. zonastab=Urbano sectoresta=Oficial calendario=A num_inst_zona=alto 222208 ==>
      lengmate=Minimo 206446    conf:(0.93)
   8. genero=Femenino zonastab=Urbano sectoresta=Oficial 192332 ==> lengmate=Minimo
      178092    conf:(0.93)
   9. genero=Femenino zonastab=Urbano sectoresta=Oficial calendario=A 192324 ==>
      lengmate=Minimo 178084    conf:(0.93)
  10. zonastab=Urbano sectoresta=Oficial calendario=A 396605 ==> lengmate=Minimo
      367087    conf:(0.93)
  11. zonastab=Urbano sectoresta=Oficial calendario=A num_estu_zona=medio 252023
      ==> lengmate=Minimo 233228    conf:(0.93)
  12. genero=Masculino zonastab=Urbano sectoresta=Oficial calendario=A 190319 ==>
      lengmate=Minimo 176007    conf:(0.92)
  13. genero=Masculino zonastab=Urbano sectoresta=Oficial 190330 ==>
      lengmate=Minimo 176016    conf:(0.92)
  14. jornada=M zonastab=Urbano sectoresta=Oficial calendario=A 236680 ==>
      lengmate=Minimo 217447    conf:(0.92)
  15. jornada=M zonastab=Urbano sectoresta=Oficial 236704 ==> lengmate=Minimo
      217468    conf:(0.92)
  16. zonastab=Urbano nivelsocioeconomico=bajo sectoresta=Oficial calendario=A
      204944 ==> lengmate=Minimo 188270    conf:(0.92)
  17. zonastab=Urbano nivelsocioeconomico=bajo sectoresta=Oficial 204968 ==>
      lengmate=Minimo 188291    conf:(0.92)
  18. zonastab=Urbano calendario=A num_inst_zona=alto 277873 ==> lengmate=Minimo
      254295    conf:(0.92)
  19. zonastab=Urbano num_inst_zona=alto 283883 ==> lengmate=Minimo 259514
      conf:(0.91)
  20. zonastab=Urbano nivelsocioeconomico=bajo 221287 ==> lengmate=Minimo 202190
      conf:(0.91)
  21. tipoent=no certificadas num_inst_zona=alto 189035 ==> lengmate=Minimo 172590
      conf:(0.91)
  22. zonastab=Urbano tipoent=no certificadas calendario=A 314337 ==>
      lengmate=Minimo 286609    conf:(0.91)
  23. zonastab=Urbano tipoent=no certificadas 321459 ==> lengmate=Minimo 292873
      conf:(0.91)
  24. zonastab=Urbano tipoent=no certificadas calendario=A num_estu_zona=medio
      240672 ==> lengmate=Minimo 219178    conf:(0.91)
```

Figura 13 Mejores reglas generadas con Apriori con el conjunto de datos de Lenguaje y Matemáticas.

- ✓ **Competencias Genéricas de Matemáticas y Ciencias Naturales (Matecien)**

De igual manera, se obtuvieron reglas de asociación que relacionan solo los factores para la prueba de Matemáticas Ciencias Naturales. Se estableció diferentes valores para los parámetros del soporte y la confianza pero los que generaron las reglas fuertes tienen como mínima confianza el 91% (0.91), un soporte mínimo superior de 1.0, un soporte mínimo inferior de 0.25, un Δ de 0.5 y un número de reglas a generar de 10.000. También se filtraron las reglas para obtener solo aquellas, donde el atributo Matecien se encuentre como consecuente de la regla. Las reglas fuertes fueron aquellas con un soporte mínimo del 25% (0.25). Los parámetros de ejecución del algoritmo Apriori se presentan en la figura 14 y las mejores 24 reglas generadas con una confianza del 91%, en la figura 15.

```
=== Run information ===
Scheme:       weka.associations.Apriori -N 100 -T 0 -C 0.91 -D 0.05 -U 1.0 -M
     0.25 -S -1.0 -A -c 5
Relation:     minable_mate_cienl-weka.filters.unsupervised.attribute.Remove R3-
     4,7-11,15,20
Instances:    444318
Attributes:   11
              jornada
              genero
              zonastab
              nivelsocioeconomico
              matecien
              tipoent
              sectoresta
              calendario
              region
              num_estu_zona
              num_inst_zona
=== Associator model (full training set) ===

Apriori
=======
Minimum support: 0.25 (111080 instances)
Minimum metric <confidence>: 0.91
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 44
Size of set of large itemsets L(3): 51
Size of set of large itemsets L(4): 17
```

Figura 14 Parámetros de ejecución del algoritmo Apriori para el repositorio Matemáticas y Ciencias Naturales

```
 Best rules found:
1.   zonastab=Urbano tipoent=no certificadas sectoresta=Oficial
     calendario=A 145298 ==> matecien=Minimo 136933    conf:(0.94)
2.   tipoent=no certificadas sectoresta=Oficial calendario=A
     num_estu_zona=medio 122089 ==> matecien=Minimo 114351    conf:(0.94)
3.   tipoent=no certificadas sectoresta=Oficial num_estu_zona=medio 122100
     ==> matecien=Minimo 114361    conf:(0.94)
4.   zonastab=Urbano sectoresta=Oficial calendario=A num_inst_zona=alto
     146628 ==> matecien=Minimo 136841    conf:(0.93)
5.   zonastab=Urbano sectoresta=Oficial num_inst_zona=alto 146639 ==>
     matecien=Minimo 136851    conf:(0.93)
6.   tipoent=no certificadas sectoresta=Oficial calendario=A 165632 ==>
     matecien=Minimo 154482    conf:(0.93)
7.   zonastab=Urbano sectoresta=Oficial num_estu_zona=medio 168222 ==>
     matecien=Minimo 156163    conf:(0.93)
8.   genero=Femenino zonastab=Urbano sectoresta=Oficial 127280 ==>
     matecien=Minimo 118125    conf:(0.93)
9.   zonastab=Urbano sectoresta=Oficial 263315 ==> matecien=Minimo 244202
     conf:(0.93)
10.  zonastab=Urbano sectoresta=Oficial calendario=A 263304 ==>
     matecien=Minimo 244192    conf:(0.93)
11.  genero=Masculino zonastab=Urbano sectoresta=Oficial 125636 ==>
     matecien=Minimo 116439    conf:(0.93)
12.  genero=Masculino zonastab=Urbano sectoresta=Oficial calendario=A
     125629 ==> matecien=Minimo 116432    conf:(0.93)
13.  jornada=M zonastab=Urbano sectoresta=Oficial calendario=A 152748 ==>
     matecien=Minimo 141448    conf:(0.93)
14.  jornada=M zonastab=Urbano sectoresta=Oficial 152759 ==>
     matecien=Minimo 141458    conf:(0.93)
15.  zonastab=Urbano nivelsocioeconomico=bajo sectoresta=Oficial
     calendario=A 148465 ==> matecien=Minimo 136510    conf:(0.92)
16.  zonastab=Urbano num_inst_zona=alto 187346 ==> matecien=Minimo 171745
     conf:(0.92)
17.  zonastab=Urbano nivelsocioeconomico=bajo calendario=A 159947 ==>
     matecien=Minimo 146309    conf:(0.91)
18.  zonastab=Urbano nivelsocioeconomico=bajo 160321 ==> matecien=Minimo
     146635    conf:(0.91)
19.  tipoent=no certificadas calendario=A num_inst_zona=alto 121700 ==>
     matecien=Minimo 111295    conf:(0.91)
20.  zonastab=Urbano tipoent=no certificadas calendario=A 211347 ==>
     matecien=Minimo 193076    conf:(0.91)
21.  zonastab=Urbano tipoent=no certificadas 215203 ==> matecien=Minimo
     196480    conf:(0.91)
22.  tipoent=no certificadas calendario=A num_estu_zona=medio 174171 ==>
     matecien=Minimo 158800    conf:(0.91)
23.  tipoent=no certificadas num_estu_zona=medio 178297 ==>
     matecien=Minimo 162507    conf:(0.91)
24.  nivelsocioeconomico=medio tipoent=no certificadas num_estu_zona=medio
     124746 ==> matecien=Minimo 113570    conf:(0.91)
```

Figura 15 Mejores reglas generadas con Apriori con el conjunto de datos de Matemáticas y Ciencias Naturales.

- ✓ **Competencias Genéricas de Matemáticas y Competencias Ciudadanas (Matecomp)**

De igual manera, se obtuvieron reglas de asociación que relacionan solo los factores para la prueba de Matemáticas Competencias Ciudadanas. Se estableció diferentes valores para los parámetros del soporte y la confianza pero los que generaron las reglas fuertes tienen como mínima confianza el 93% (0.93), un soporte mínimo superior de 1.0, un soporte mínimo inferior de 0.25, un Δ de 0.5 y un número de reglas a generar de 10.000. También se filtraron las reglas para obtener solo aquellas, donde el atributo Matecomp se encuentre como consecuente de la regla. Las reglas fuertes fueron aquellas con un soporte mínimo del 25% (0.25). Los parámetros de ejecución del algoritmo Apriori se presentan en la figura 16 y las mejores 24 reglas generadas con una confianza del 93%, en la figura 17.

```
=== Run information ===
Scheme:       weka.associations.Apriori -N 100 -T 0 -C 0.93 -D 0.05 -U 1.0 -M 0.25 -S -1.0
-A -c 5
Relation:     minable_mate_compl-weka.filters.unsupervised.attribute.Remove-R3-4,6,8-
12,16,21
Instances:    220248
Attributes:   11
               jornada
               genero
               zonastab
               nivelsocioeconomico
               matecomp
               tipoent
               sectoresta
               calendario
               region
               num_estu_zona
               num_inst_zona
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.25 (55062 instances)
Minimum metric <confidence>: 0.93
Number of cycles performed: 15
Generated sets of large itemsets:
Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 48
Size of set of large itemsets L(3): 57
Size of set of large itemsets L(4): 22
```

Figura 16 Parámetros de ejecución del algoritmo Apriori para el repositorio Matemáticas y Competencias Ciudadanas.

```
Best rules found:

1.   zonastab=Urbano tipoent=no certificadas sectoresta=Oficial 70825 ==>
     matecomp=Minimo 67320    conf:(0.95)
2.   zonastab=Urbano tipoent=no certificadas sectoresta=Oficial calendario=A
     70809 ==> matecomp=Minimo 67304    conf:(0.95)
3.   zonastab=Urbano nivelsocioeconomico=medio sectoresta=Oficial 74556 ==>
     matecomp=Minimo 70579    conf:(0.95)
4.   zonastab=Urbano nivelsocioeconomico=medio sectoresta=Oficial calendario=A
     74556 ==> matecomp=Minimo 70579    conf:(0.95)
5.   nivelsocioeconomico=medio sectoresta=Oficial num_estu_zona=medio 58597 ==>
     matecomp=Minimo 55417    conf:(0.95)
6.   nivelsocioeconomico=medio sectoresta=Oficial calendario=A
     num_estu_zona=medio 58597 ==> matecomp=Minimo 55417    conf:(0.95)
7.   nivelsocioeconomico=medio sectoresta=Oficial 76505 ==> matecomp=Minimo
     72335    conf:(0.95)
8.   nivelsocioeconomico=medio sectoresta=Oficial calendario=A 76505 ==>
     matecomp=Minimo 72335    conf:(0.95)
9.   genero=Femenino zonastab=Urbano sectoresta=Oficial 63065 ==>
     matecomp=Minimo 59468    conf:(0.94)
10.  genero=Femenino zonastab=Urbano sectoresta=Oficial calendario=A 63063 ==>
     matecomp=Minimo 59466    conf:(0.94)
11.  zonastab=Urbano sectoresta=Oficial num_inst_zona=alto 73543 ==>
     matecomp=Minimo 69317    conf:(0.94)
12.  zonastab=Urbano sectoresta=Oficial calendario=A num_inst_zona=alto 73527
     ==> matecomp=Minimo 69301    conf:(0.94)
13.  tipoent=no certificadas sectoresta=Oficial calendario=A 80936 ==>
     matecomp=Minimo 76260    conf:(0.94)
14.  tipoent=no certificadas sectoresta=Oficial 80978 ==> matecomp=Minimo 76292
     conf:(0.94)
15.  zonastab=Urbano sectoresta=Oficial 130100 ==> matecomp=Minimo 122545
     conf:(0.94)
16.  zonastab=Urbano sectoresta=Oficial calendario=A 130084 ==> matecomp=Minimo
     122529    conf:(0.94)
17.  tipoent=no certificadas sectoresta=Oficial calendario=A num_estu_zona=medio
     59337 ==> matecomp=Minimo 55849    conf:(0.94)
18.  genero=Masculino zonastab=Urbano sectoresta=Oficial 62479 ==>
     matecomp=Minimo 58780    conf:(0.94)
19.  genero=Masculino zonastab=Urbano sectoresta=Oficial calendario=A 62470 ==>
     matecomp=Minimo 58771    conf:(0.94)
20.  zonastab=Urbano sectoresta=Oficial num_estu_zona=medio 82183 ==>
     matecomp=Minimo 77265    conf:(0.94)
21.  zonastab=Urbano sectoresta=Oficial calendario=A num_estu_zona=medio 82167
     ==> matecomp=Minimo 77249    conf:(0.94)
22.  jornada=M zonastab=Urbano sectoresta=Oficial 82004 ==> matecomp=Minimo
     76944    conf:(0.94)
23.  jornada=M zonastab=Urbano sectoresta=Oficial calendario=A 81988 ==>
     matecomp=Minimo 76928    conf:(0.94)
24.  zonastab=Urbano tipoent=municipio si sectoresta=Oficial calendario=A 59275
     ==> matecomp=Minimo 55225    conf:(0.93)
```

Figura 17 Mejores reglas generadas con Apriori con el conjunto de datos de Matemáticas y Competencias Ciudadanas.

3.4.2 Tarea de Clustering

Para esta tarea se cuenta con un repositorio de datos minable del cual se han generado cinco tablas, una por cada par de competencias presentadas en las Pruebas Saber 5. En este caso las variables que se tienen en cuenta para aplicar el algoritmo están descritas en la tabla 39. Como en las tareas anteriores lo importante es encontrar *clusters* donde haya al menos un grupo de cada nivel de desempeño (Insuficiente, Mínimo, Satisfactorio y Avanzado). Para la generación de los *clusters* se usó la herramienta de minería de datos Weka, que contiene la implementación del algoritmo de aprendizaje simple k-means. Se configuró el número de clusters según la distribución porcentual de los datos, con una semilla por defecto de 10. Para evaluar los resultados del agrupamiento se utilizó el propio conjunto de entrenamiento Use training set.

3.4.2.1. Clusters Generados con el Algoritmo Simple k-means

Competencias Genéricas de Lenguaje y Ciencias Naturales (Lengcien)

Para encontrar grupos que relacionen los factores asociados al rendimiento en las Pruebas Saber 5 se utilizó los resultados obtenidos para el repositorio de datos correspondiente a estudiantes que presentaron la prueba que contenía las competencias de Lenguaje y Ciencias Naturales. Se configuró el número de clusters en 3 según la distribución porcentual de los datos, con una semilla por defecto de 10. En la figura 18 se muestra la configuración usada para el algoritmo simple k-menas y en la figura 19 se pueden observar los resultados clasificando en el *cluster 0* a 226509 (51%) del total de estudiantes; en el *cluster 1* se agrupan 123310 (28%) y en el *cluster 2* se encuentran 95880 (22%). Los atributos que los determinan corresponden a las características del centroide de cada grupo y se toma la moda para los atributos nominales (Hernández, Ramirez & Ferri, 2005).

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-
pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     minable_leng_cien1-weka.filters.unsupervised.attribute.Remove-R20
Instances:    445699
Attributes:   18
              jornada
              genero
              sector
              zona
              zonaestab
              nivelsocioeconomico
              discapacidad
              leng_copietas
              cien_copietas
              lenguaje
              ciencias
              tipoent
              sectoresta
              tipoestablecimiento
```

Figura 18 Configuración del algoritmo simple k-means para las competencias de Lenguaje y Ciencias naturales.

k-Means

Number of iterations: 5

Final cluster centroids:

		Cluster #		
Attribute	Full Data	0	1	2
	(445699.0)	(226509.0)	(123310.0)	(95880.0)
jornada	M	M	M	M
genero	Masculino	Masculino	Femenino	Femenino
sector	Oficial	Oficial	Oficial	Oficial
zona	Urbano	Urbano	Urbano	Rural
zonaestab	Urbano	Urbano	Urbano	Rural
nivelsocioeconomico	bajo	medio	medio	bajo
discapacidad	SIN DISCAPACIDAD	SIN DISCAPACIDAD	SIN DISCAPACIDAD	SIN DISCAPACIDAD
leng_copietas	No copia	No copia	No copia	No copia
cien_copietas	No copia	No copia	No copia	No copia
lenguaje	Insuficiente	Insuficiente	Minimo	Insuficiente
ciencias	Insuficiente	Insuficiente	Minimo	Insuficiente
tipoent	no certificadas	no certificadas	no certificadas	municipio si
sectoresta	Oficial	Oficial	Oficial	Oficial
tipoestablecimiento	Oficial urbano	Oficial urbano	Oficial urbano	Oficial rural
calendario	A	A	A	A
region	caribe	caribe	bogota	caribe
num_estu_zona	medio	medio	medio	medio
num_inst_zona	alto	alto	bajo	alto

Clustered Instances

0	226509	(51%)
1	123310	(28%)
2	95880	(22%)

Figura 19 Descripción de clusters para las competencias de Lenguaje y Ciencias Naturales.

- ✓ **Competencias Genéricas de Lenguaje y Competencias Ciudadanas (Lengcomp)**

Para encontrar grupos que relacionen los factores asociados al rendimiento en las Pruebas Saber 5 en cuanto a la prueba de Lenguaje y Competencias Ciudadanas, se configuró el número de clusters en 4 según la distribución porcentual de los datos, con una semilla por defecto de 10. En la figura 20 se muestra la configuración usada para el algoritmo simple k-menas. En la figura 21 se pueden observar los resultados del algoritmo, clasificando en el *cluster 0* a 89293 (41%) del total de estudiantes; en el *cluster 1* se agrupan 25008 (11%); en el *cluster 2* se encuentran 17071 (8%) y en el *cluster 3* se concentran 17071 (40%). Los atributos que los determinan corresponden a las características del centroide de cada grupo y se toma la moda para los atributos nominales (Hernández, Ramirez & Ferri, 2005).

```
===Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -
periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 4 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     minable_leng_compl-weka.filters.unsupervised.attribute.Remove-R20
Instances:    219748
Attributes:   18
              jornada
              genero
              sector
              zona
              zonastab
              nivelsocioeconomico
              discapacidad
              leng_copietas
              comp_copietas
              lenguaje
              competencias
              tipoent
              sectoresta
              tipoestablecimiento
              calendario
              region
              num_estu_zona
              num_inst_zona
Test mode:    evaluate on training data
```

Figura 20 Configuración del algoritmo simple k-means para las competencias de Lenguaje y Competencias Ciudadanas

.

```
k-Means
======

Number of iterations: 4
Final cluster centroids:
                                                Cluster#
Attribute                 Full Data                 0                 1                 2                 3
                          (219748.0)        (89293.0)         (25008.0)         (17071.0)         (88376.0)
=============================================================================================================
jornada                           M                 M                 M                 C                 M
genero                    Masculino         Femenino         Masculino         Masculino         Masculino
sector                      Oficial           Oficial       No Oficial       No Oficial           Oficial
zona                          Urbano            Urbano           Urbano           Urbano            Urbano
zonastab                      Urbano            Urbano           Urbano           Urbano            Urbano
nivelsocioeconomico           medio             medio            medio            medio             bajo
discapacidad      SIN DISCAPACIDAD  SIN DISCAPACIDAD  SIN DISCAPACIDAD  SIN DISCAPACIDAD  SIN DISCAPACIDAD
leng_copietas          No copia          No copia          No copia          No copia          No copia
comp_copietas          No copia          No copia          No copia          No copia          No copia
lenguaje                 Minimo            Minimo           Minimo            Minimo            Minimo
competencias             Minimo            Minimo           Minimo            Minimo            Minimo
lengcomp                 Minimo            Minimo           Minimo            Minimo            Minimo
tipoent          no certificadas   no certificadas   no certificadas   no certificadas     municipio si
sectoresta                Oficial           Oficial       No Oficial       No Oficial           Oficial
tipoestablecimiento  Oficial urbano   Oficial urbano       No Oficial       No Oficial   Oficial urbano
calendario                    A                 A                 A                 A                 A
region                   caribe      eje cafetero           caribe           bogota            caribe
num_estu_zona             medio             medio            medio            medio             medio
num_inst_zona              alto              alto             alto             bajo              alto


=== Model and evaluation on training set ===

Clustered Instances

0       89293 ( 41%)
1       25008 ( 11%)
2       17071 (  8%)
3       88376 ( 40%)
```

Figura 21 Descripción de clusters para las competencias de Lenguaje y Competencias Ciudadanas.

- ✓ **Competencias Genéricas de Lenguaje y Matemáticas (Lengmate)**

Para encontrar grupos que relacionen los factores asociados al rendimiento en las Pruebas Saber 5 en cuanto a la prueba de Lenguaje y Matemáticas, se configuró el número de clusters en 3 según la distribución porcentual de los datos, con una semilla por defecto de 10. En la figura 22 se muestra la configuración usada para el algoritmo simple k-menas. En la figura 23 se pueden observar los resultados del algoritmo, clasificando en el *cluster 0* a 136263 (20%) del total de estudiantes; en el *cluster 1* se agrupan 209072 (31%) y en el *cluster 2* se encuentran 325717 (49%). Los atributos que los determinan corresponden a las características del centroide de cada grupo y se toma la moda para los atributos nominales (Hernández, Ramirez & Ferri, 2005).

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -
periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     minable_leng_matel-weka.filters.unsupervised.attribute.Remove-
R12,20
Instances:    671052
Attributes:   18
              jornada
              genero
              sector
              zona
              zonastab
              nivelsocioeconomico
              discapacidad
              leng_copietas
              mate_copietas
              lenguaje
              matematicas
              tipoent
              sectoresta
              tipoestablecimiento
              calendario
              region
              num_estu_zona
              num_inst_zona
Test mode:    evaluate on training data
```

Figura 22 Configuración del algoritmo simple k-means para las competencias de Lenguaje y Matemáticas.

```
k-Means
======

Number of iterations: 3
Final cluster centroids:
                                          Cluster#
Attribute                    Full Data           0                1                2
                            (671052.0)  (136263.0)       (209072.0)       (325717.0)
=============================================================================================
jornada                              M           C                M                M
genero                       Masculino   Masculino        Femenino        Masculino
sector                         Oficial  No Oficial         Oficial          Oficial
zona                             Urbano      Urbano            Rural           Urbano
zonastab                         Urbano      Urbano            Rural           Urbano
nivelsocioeconomico                bajo       medio             bajo            medio
discapacidad          SIN DISCAPACIDAD SIN DISCAPACIDAD SIN DISCAPACIDAD SIN DISCAPACIDAD
leng_copietas              No copia    No copia         No copia         No copia
mate_copietas              No copia    No copia         No copia         No copia
lenguaje                     Minimo  Insuficiente     Insuficiente           Minimo
matematicas                  Minimo  Insuficiente     Insuficiente           Minimo
tipoent               no certificadas no certificadas municipio si  no certificadas
sectoresta                     Oficial  No Oficial         Oficial          Oficial
tipoestablecimiento     Oficial urbano  No Oficial    Oficial rural   Oficial urbano
calendario                         A           A                A                A
region                        caribe  centro orient         caribe           caribe
num_estu_zona                  medio       medio            medio            medio
num_inst_zona                  alto        medio             alto             alto

Clustered Instances

0      136263 ( 20%)
1      209072 ( 31%)
2      325717 ( 49%)
```

Figura 23 Descripción de clusters para las competencias de Lenguaje y Matemáticas

- ✓ **Competencias genéricas de Matemáticas y Ciencias Naturales (Matecien)**

Para encontrar grupos que relacionen los factores asociados al rendimiento en las Pruebas Saber 5 en cuanto a la prueba de Matemáticas y Ciencias Naturales, se configuró el número de clusters en 3 según la distribución porcentual de los datos, con una semilla por defecto de 10. En la figura 24 se muestra la configuración usada para el algoritmo simple k-menas. En la figura 25 se pueden observar los resultados del algoritmo, clasificando en el cluster 0 a 235607 (53%) del total de estudiantes; en el cluster 1 se agrupan 118793 (27%) y en el cluster 2 se encuentran 89918 (20%). Los atributos que los determinan corresponden a las características del centroide de cada grupo y se toma la moda para los atributos nominales (Hernández, Ramirez & Ferri, 2005).

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -
periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     minable_mate_cienl-weka.filters.unsupervised.attribute.Remove-
R12,20
Instances:    444318
Attributes:   18
              jornada
              genero
              sector
              zona
              zonastab
              nivelsocioeconomico
              discapacidad
              mate_copietas
              cien_copietas
              matematicas
              ciencias
              tipoent
              sectoresta
              tipoestablecimiento
              calendario
              region
              num_estu_zona
              num_inst_zona
Test mode:    evaluate on training data
```

Figura 24 Configuración del algoritmo simple k-means para las competencias de Matemáticas y Ciencias Naturales.

```
k-Means
======

Number of iterations: 4
Final cluster centroids:
                                                Cluster#
Attribute                    Full Data          0                1                2
                            (444318.0)       (235607.0)       (118793.0)       (89918.0)
=============================================================================================
jornada                              M                M                M                M
genero                        Masculino        Femenino        Masculino        Masculino
sector                          Oficial          Oficial          Oficial          Oficial
zona                              Urbano           Urbano           Urbano            Rural
zonastab                          Urbano           Urbano           Urbano            Rural
nivelsocioeconomico                bajo            medio             bajo             bajo
discapacidad          SIN DISCAPACIDAD SIN DISCAPACIDAD SIN DISCAPACIDAD SIN DISCAPACIDAD
mate_copietas               No copia         No copia         No copia         No copia
cien_copietas               No copia         No copia         No copia         No copia
matematicas              Insuficiente     Insuficiente           Minimo     Insuficiente
ciencias                 Insuficiente     Insuficiente           Minimo     Insuficiente
tipoent                no certificadas  no certificadas  no certificadas     municipio si
sectoresta                      Oficial          Oficial          Oficial          Oficial
tipoestablecimiento     Oficial urbano   Oficial urbano   Oficial urbano    Oficial rural
calendario                          A                A                A                A
region                         caribe           caribe           caribe           caribe
num_estu_zona                  medio            medio            medio            medio
num_inst_zona                   alto             alto             alto             alto


Clustered Instances

0      235607 ( 53%)
1      118793 ( 27%)
2       89918 ( 20%)
```

Figura 25 . Descripción de clusters para las competencias de Matemáticas y Ciencias Naturales.

- ✓ **Competencias Genéricas de Matemáticas y Competencias Ciudadanas (Matecomp)**

Para encontrar grupos que relacionen los factores asociados al rendimiento en las Pruebas Saber 5 en cuanto a la prueba de Matemáticas y Competencias Ciudadanas, se configuró el número de clusters en 3 según la distribución porcentual de los datos, con una semilla por defecto de 10. En la figura 26 se muestra la configuración usada para el algoritmo simple k-menas. En la figura 27 se pueden observar los resultados del algoritmo, clasificando en el *cluster 0* a 112734 (51%) del total de estudiantes; en el *cluster 1* se agrupan 50779 (23%) y en el *cluster 2* se encuentran 56735 (26%). Los atributos que los determinan corresponden a las características del centroide de cada grupo y se toma la moda para los atributos nominales (Hernández, Ramirez & Ferri, 2005).

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning
10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-
last" -I 500 -num-slots 1 -S 10
Relation:     minable_mate_compl-weka.filters.unsupervised.attribute.Remove-R6,13,21
Instances:    220248
Attributes:   18
              jornada
              genero
              sector
              zona
              zonastab
              nivelsocioeconomico
              discapacidad
              mate_copietas
              comp_copietas
              matematicas
              competencias
              tipoent
              sectoresta
              tipoestablecimiento
              cal_establecimiento
              region
              num_estu_zona
              num_inst_zona
Test mode:    evaluate on training data
```

Figura 26 Configuración del algoritmo simple k-means para las competencias de Matemáticas y Competencias Ciudadanas.

```
k-Means
======

Number of iterations: 4
Final cluster centroids:
                                        Cluster#
Attribute                 Full Data             0                1                2
                         (220248.0)       (112734.0)       (50779.0)        (56735.0)
=============================================================================================
jornada                           M                M                M                T
genero                     Masculino        Masculino         Femenino         Femenino
sector                       Oficial          Oficial          Oficial          Oficial
zona                          Urbano           Urbano            Rural           Urbano
zonastab                      Urbano           Urbano            Rural           Urbano
nivelsocioeconomico            medio            medio             bajo            medio
discapacidad          SIN DISCAPACIDAD SIN DISCAPACIDAD SIN DISCAPACIDAD SIN DISCAPACIDAD
mate_copietas              No copia         No copia         No copia         No copia
comp_copietas              No copia         No copia         No copia         No copia
matematicas                Minimo           Minimo           Minimo           Minimo
competencias               Minimo           Minimo           Minimo           Minimo
tipoent                no certificadas     municipio si     municipio si  no certificadas
sectoresta                  Oficial          Oficial          Oficial          Oficial
tipoestablecimiento     Oficial urbano   Oficial urbano    Oficial rural   Oficial urbano
cal_establecimiento                A                A                A                A
region                        caribe           caribe           caribe           bogota
num_estu_zona                 medio            medio            medio            medio
num_inst_zona                  alto             alto             alto             bajo


Clustered Instances

0      112734 ( 51%)
1       50779 ( 23%)
2       56735 ( 26%)
```

Figura 27 Descripción de clusters para las competencias de Matemáticas y Competencias Ciudadanas.

3.5 Evaluación

En la siguiente sección se evalúan e interpretan los resultados obtenidos con los datos de estudiantes que presentaros las Pruebas Saber 5 en el periodo comprendido entre los años 2014 y 2016 almacenados en las cinco bases de datos minables: Lengcien, Lengcomp, Lengmate, Matecien, Matecomp aplicando las tareas de asociación y clusteirng.

En esta etapa se evalúa la calidad de los modelos obtenidos y se interpretan los patrones descubiertos con el fin de consolidar el conocimiento descubierto e incorporarlo al existente para posteriores acciones o para confrontarlo con conocimiento previamente descubierto. Esta etapa incluye la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles en términos que sean entendibles para el usuario.

Las técnicas de minería de datos producen modelos que explican de manera general los datos y su comprensibilidad al ser un factor subjetivo, depende en gran modo de la experiencia y conocimiento de los usuarios de los modelos (Hernández, Ramirez & Ferri, 2005). Los sistemas de reglas, como los utilizados en esta investigación, son considerados como una de las representaciones que permiten comprender más fácilmente el comportamiento de un modelo, sin embargo, por el número de reglas, el número de atributos y las métricas de confiabilidad utilizadas, se hace necesaria su interpretación para facilitarles la comprensibilidad a los usuarios. Por otra parte, los diferentes modelos pueden producir patrones similares si se trabajan con repositorios de datos comunes.

3.5.1 Modelos de Asociación

- ✓ **Competencias Genéricas de Lenguaje y Ciencias Naturales (Lengcien)**

En la figura 8, se muestran los resultados obtenidos mediante el algoritmo A priori, para la generación del modelo del peso promedio entre las competencias de Lenguaje y Ciencias naturales

(lengcien); se puede observar que el modelo generado tiene un soporte del 0.2, con lo cual ha predicho un total de 445699 instancias de forma correcta, con un mínimo de confianza previamente establecido de 0.92. El algoritmo Apriori ha utilizado un total de 16 ciclos para la generación del modelo. Además, en la figura 9, se visualiza que las 20 primeras reglas, son fuertes ya que poseen un nivel de confianza mayor que 0.92.

A continuación se interpretan las reglas que se consideran como las más relevantes puesto que presentan un nivel de confianza mayor a la establecida anteriormente.

Regla 1: el 94% de los estudiantes que presentaron la prueba de Lenguaje y Ciencias Naturales que obtuvieron un nivel mínimo, pertenecen a establecimientos educativos con jornada de la tarde y son del sector oficial. El 21% de todos los estudiantes que presentaron la prueba cumplen con este patrón.

Regla 5: el 94% de los estudiantes que presentaron la prueba de Lenguaje y Ciencias Naturales que obtuvieron un nivel mínimo, pertenecen a establecimientos educativos de la zona urbana, son del sector oficial y el número de instituciones por zona es alto, es decir es mayor a 4000 instituciones. El 31% de todos los estudiantes que presentaron la prueba cumplen con este patrón.

Regla 8: el 94% de los estudiantes que presentaron la prueba de Lenguaje y Ciencias Naturales que obtuvieron un nivel mínimo, pertenecen a establecimientos educativos de la zona urbana, están adscritas a entidades territoriales no certificadas, son del sector oficial y el número de estudiantes por zona es medio, es decir que hay entre 200.000 y 300.000 estudiantes. El 23% de todos los estudiantes que presentaron la prueba cumplen con este patrón.

Regla 16: el 93% de los estudiantes que presentaron la prueba de Lenguaje y Ciencias Naturales que obtuvieron un nivel mínimo, pertenecen a establecimientos educativos de la zona

urbana, su nivel socioeconómico en bajo, es decir están entre los estratos 1 y 2, tienen calendario A y el número de instituciones por zona mayor de 4000. El 21% de todos los estudiantes que presentaron la prueba cumplen con este patrón.

Regla 21*:* el 92% de los estudiantes que presentaron la prueba de Lenguaje y Ciencias Naturales que obtuvieron un nivel mínimo, son de género masculino, pertenecen a establecimientos educativos de la zona urbana, están en el sector oficial y tienen calendario A. El 26% de todos los estudiantes que presentaron la prueba cumplen con este patrón.

Regla 24: el 92% de los estudiantes que presentaron la prueba de Lenguaje y Ciencias Naturales que obtuvieron un nivel mínimo, están en jornada de la mañana, pertenecen a establecimientos educativos de la zona urbana, están en el sector oficial y tienen calendario A. El 32% de todos los estudiantes que presentaron la prueba cumplen con este patrón.

- ✓ **Competencias Genéricas de Lenguaje y Competencias Ciudadanas (Lengcomp)**

En la figura10, se muestran los resultados obtenidos mediante el algoritmo A priori, para la generación del modelo del peso promedio entre las competencias de Lenguaje y Competencias Ciudadanas (lengcomp); se puede observar que el modelo generado tiene un soporte del 0.25, con lo cual ha predicho un total de 219748 instancias de forma correcta, con un mínimo de confianza previamente establecido de 0.96. El algoritmo Apriori ha utilizado un total de 15 ciclos para la generación del modelo. Además, en la figura 11, se visualiza que las 9 primeras reglas, son las fuertes ya que poseen un nivel de confianza mayor que 0.96.

A continuación se interpretan las reglas que se consideran como las más relevantes de las 24 generadas por el modelo descrito.

Regla 5**:** el 97% de los estudiantes que presentaron la prueba de Lenguaje y Competencias Ciudadanas que obtuvieron un nivel mínimo de desempeño, pertenecen a establecimientos

educativos de nivel socioeconómico medio, son del sector oficial, tienen calendario A y el número de estudiantes por zona están entre los 200.000 y 300.000. El 26% de todos los estudiantes que presentaron la prueba cumplen con este patrón.

Regla 12: el 96% de los estudiantes que presentaron las prueba de Lenguaje y Competencias ciudadanas que obtuvieron un nivel mínimo de desempeño, pertenecen a establecimientos educativos de la zona urbana, su nivel socioeconómico es medio, están adscritos a entidades territoriales no certificadas y tienen calendario A. El 34% de todos los estudiantes que presentaron esta prueba cumplen con este patrón.

Regla 19: el 96% de los estudiantes que presentaron la prueba de Lenguaje y Competencias ciudadanas que obtuvieron un nivel mínimo de desempeño, son de género femenino, están matriculados a establecimientos educativos de la zona urbana, del sector oficial, con calendario A. El 28% de todos los estudiantes que presentaron esta prueba cumplen con este patrón.

Regla 22: el 96% de los estudiantes que presentaron la prueba de Lenguaje y Competencias ciudadanas que obtuvieron un nivel mínimo de desempeño, están en establecimientos educativos de jornada de la mañana, están ubicados en la zona urbana y son de tipo oficial. El 36% de todos los estudiantes que presentaron esta prueba cumplen con este patrón.

- ✓ **Competencias Genéricas de Lenguaje y Matemáticas (Lengmate)**

En la figura 12, se muestran los resultados obtenidos mediante el algoritmo A priori, para la generación del modelo del peso promedio entre las competencias de Lenguaje y Matemáticas (lengmate); se puede observar que el modelo generado tiene un soporte del 0.25, con lo cual ha predicho un total de 671052 instancias de forma correcta, con un mínimo de confianza previamente establecido de 0.91. El algoritmo Apriori ha utilizado un total de 15 ciclos para la generación del

modelo. Además, en la figura 13, se visualiza que las 18 primeras reglas, son las fuertes ya que poseen un nivel de confianza mayor que 0.91.

A continuación se interpretan las reglas que se consideran como las más relevantes de las 24 generadas por el modelo descrito.

Regla 1: el 94% de los estudiantes que presentaron la prueba de Lenguaje y Matemáticas que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos de la zona urbana, el tipo de entidad territorial es no certificada, son del sector oficial y tienen calendario A. El 30% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 5: el 93% de los estudiantes que presentaron la prueba de Lenguaje y Matemáticas que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos de entidades territoriales no certificadas, son del sector oficial y tienen calendario A. El 25% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 9: el 93% de los estudiantes que presentaron la prueba de Lenguaje y Matemáticas que obtuvieron un nivel de desempeño mínimo, son de género femenino, pertenecen a establecimientos educativos de la zona urbana, son del sector oficial y tienen calendario A. El 26% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 13: el 92% de los estudiantes que presentaron la prueba de Lenguaje y Matemáticas que obtuvieron un nivel de desempeño mínimo, son de género masculino, pertenecen a establecimientos educativos de la zona urbana y son del sector oficial. El 26% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 14: el 92% de los estudiantes que presentaron la prueba de Lenguaje y Matemáticas que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos con

jornada de la mañana, de la zona urbana, son del sector oficial y tienen calendario A. El 28% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón

Regla 16: el 92% de los estudiantes que presentaron la prueba de Lenguaje y Matemáticas que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos de la zona urbana, con nivel socioeconómico bajo, es decir están dentro de los estratos 1 y 2, son del sector oficial y tienen calendario A. El 32% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 21: el 91% de los estudiantes que presentaron la prueba de Lenguaje y Matemáticas que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos pertenecientes a entidades territoriales no certificadas y el número de instituciones por zona es alto, es decir, que son mayores de 4000 instituciones. El 26% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

- ✓ **Competencias Genéricas de Matemáticas y Ciencias Naturales (Matecien)**

En la figura 14, se muestran los resultados obtenidos mediante el algoritmo A priori, para la generación del modelo del peso promedio entre las competencias de Matemáticas y Ciencias naturales (matecien); se puede observar que el modelo generado tiene un soporte del 0.25, con lo cual ha predicho un total de 444318 instancias de forma correcta, con un mínimo de confianza previamente establecido de 0.91. El algoritmo Apriori ha utilizado un total de 15 ciclos para la generación del modelo. Además, en la figura 15, se visualiza que las 16 primeras reglas, son las fuertes ya que poseen un nivel de confianza mayor que 0.91.

A continuación se interpretan las reglas que se consideran como las más relevantes de las 24 generadas por el modelo descrito.

Regla 2: el 94% de los estudiantes que presentaron la prueba de Matemáticas y Ciencias Naturales que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos donde el tipo de entidad territorial es no certificada, son del sector oficial, tienen calendario A y el número de estudiantes por zona están entre los 200.000 y 300.000. El 28% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 5: el 93% de los estudiantes que presentaron la prueba de Matemáticas y Ciencias Naturales que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos ubicados en la zona urbana, son del sector oficial, y el número de instituciones por zona es mayor de 4000. El 31% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 7: el 93% de los estudiantes que presentaron la prueba de Matemáticas y Ciencias Naturales que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos ubicados en la zona urbana, son del sector oficial, y el número de estudiantes por zona está entre 200.000 y 300.000. El 35% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 8: el 93% de los estudiantes que presentaron la prueba de Matemáticas y Ciencias Naturales que obtuvieron un nivel de desempeño mínimo son de género femenino pertenecen a establecimientos educativos ubicados en la zona urbana y son del sector oficial. El 27% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 12: el 93% de los estudiantes que presentaron la prueba de Matemáticas y Ciencias Naturales que obtuvieron un nivel de desempeño mínimo, son de género masculino pertenecen a establecimientos educativos ubicados en la zona urbana y son del sector oficial. El 26% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 13: el 93% de los estudiantes que presentaron la prueba de Matemáticas y Ciencias Naturales que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos con jornada de la mañana, de la zona urbana, de sector oficial y tienen calendario A. El 32% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 15: el 92% de los estudiantes que presentaron la prueba de Matemáticas y Ciencias Naturales que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos ubicados en la zona urbana, el nivel socioeconómico del establecimiento es bajo, esto quiere decir que pueden ser de estrato 1 y 2, pertenecen al sector oficial y tienen calendario A. El 31% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

- ✓ **Competencias Genéricas de Matemáticas y Competencias Ciudadanas (Matecomp)**

En la figura 16, se muestran los resultados obtenidos mediante el algoritmo A priori, para la generación del modelo del peso promedio entre las competencias de Matemáticas y Competencias ciudadanas (matecomp); se puede observar que el modelo generado tiene un soporte del 0.25, con lo cual ha predicho un total de 220248 instancias de forma correcta, con un mínimo de confianza previamente establecido de 0.93. El algoritmo Apriori ha utilizado un total de 15 ciclos para la generación del modelo. Además, en la figura 17, se visualiza que las 23 primeras reglas, son las fuertes ya que poseen un nivel de confianza mayor que 0.93.

A continuación se interpretan las reglas que se consideran como las más relevantes de las 24 generadas por el modelo descrito.

Regla 2: el 95% de los estudiantes que presentaron la prueba de Matemáticas y Competencias ciudadanas que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos ubicados en zona urbana, el tipo de entidad territorial es no certificada,

son del sector oficial y tienen calendario A. El 31% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 5: el 95% de los estudiantes que presentaron la prueba de Matemáticas y Competencias ciudadanas que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos con nivel socioeconómico medio, es decir de estratificación 3 y 4, son del sector oficial y el número de estudiantes por zona están entre los 200.000 y 300.000. El 25% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 9: el 94% de los estudiantes que presentaron la prueba de Matemáticas y Competencias ciudadanas que obtuvieron un nivel de desempeño mínimo, son de género femenino, pertenecen a establecimientos educativos ubicados en zona urbana y son del sector oficial. El 27% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 12: el 94% de los estudiantes que presentaron la prueba de Matemáticas y Competencias ciudadanas que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos ubicados en zona urbana, son del sector oficial, tienen calendario A y el número de instituciones por zona es mayor de 4000. El 31% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 19: el 94% de los estudiantes que presentaron la prueba de Matemáticas y Competencias ciudadanas que obtuvieron un nivel de desempeño mínimo, son de género masculino, pertenecen a establecimientos educativos ubicados en zona urbana, son del sector oficial y tienen calendario A. El 27% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 21: el 94% de los estudiantes que presentaron la prueba de Matemáticas y Competencias ciudadanas que obtuvieron un nivel de desempeño mínimo, pertenecen a

establecimientos educativos ubicados en zona urbana, del sector oficial, con calendario A y un número de estudiantes por zona entre los 200.000 y 300.000. El 35% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 22: el 94% de los estudiantes que presentaron la prueba de Matemáticas y Competencias ciudadanas que obtuvieron un nivel de desempeño mínimo, son de jornada de la mañana, pertenecen a establecimientos educativos ubicados en zona urbana y son del sector oficial. El 35% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

Regla 24: el 93% de los estudiantes que presentaron la prueba de Matemáticas y Competencias ciudadanas que obtuvieron un nivel de desempeño mínimo, pertenecen a establecimientos educativos ubicados en zona urbana, su tipo de entidad territorial es certificada, son del sector oficial y tienen calendario A. El 26% de todos los estudiantes que presentaron esta prueba cumplen el mismo patrón.

En la aplicación del modelo de asociación con el algoritmo Apriori, se descartan variables como: discapacidad e indicio de copia debido a que el número de atributos en un nivel de la variable supera de manera considerable a otros niveles. La variable departamento, presenta treinta y tres subcategorías que el algoritmo Apriori no tiene en cuenta al momento de generar las reglas. Y finalmente, la variable tipo de establecimiento es explicada por los atributos zona y el sector.

Los resultados obtenidos con las reglas de asociación nos permitieron conocer los factores asociados al nivel de desempeño mínimo, entre los cuales están: la zona, el sector, el nivel socioeconómico, el calendario y la entidad territorial del establecimiento educativo y el género del estudiante.

3.5.2 Modelos de Clustering

- ✓ **Competencias Genéricas de Lenguaje y Ciencias Naturales (Lengcien)**

De acuerdo a los resultados obtenidos en la figura 19, se generó 3 grupos similares en los cuatro niveles de desempeño contenidos en las competencias genéricas de Lenguaje y Ciencias Naturales, se realizaron 5 iteraciones donde se clasifica el total de estudiantes (445699) en tres clusters. En el *cluster* 0 se encuentran el 51% del total de estudiantes; en el *cluster* 1 se agrupa el 28%; en el *cluster* 2 está el 22%.

Al realizar la lectura de los resultados de cada *cluster*, se pueden obtener los siguientes patrones descriptivos:

***Cluster 0*:** El 51% de los estudiantes que presentaron la prueba de Lenguaje y Ciencias Naturales obtienen niveles de desempeño Insuficiente en las dos competencias, son de género masculino, no presentan ningún tipo de discapacidad ni indicios de copia en la prueba, están ubicados en la región Caribe, se encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Urbana, en la jornada de la mañana y calendario A, con entidad territorial no certificada, además el número instituciones en la región es superior a 4000 y el número de estudiantes por región está entre 200.000 y 300.000, y las instituciones poseen un nivel socioeconómico que oscila entre 3 y 4.

***Cluster 1*:** El 28% de los estudiantes que presentaron la prueba de Lenguaje y Ciencias Naturales obtienen niveles de desempeño Mínimo en las dos competencias, son de género femenino, no presentan ningún tipo de discapacidad ni indicios de copia en la prueba, están ubicados en la región de Bogotá, se encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Urbana, en la jornada de la mañana y calendario A, con entidad territorial no certificada, además el número instituciones en la región es inferior a 2000 y el número de estudiantes por región está entre 200.000 y 300.000, y las instituciones poseen un nivel socioeconómico que oscila entre 3 y 4.

***Cluster 2*:** El 22% de los estudiantes que presentaron la prueba de Lenguaje y Ciencias Naturales obtienen niveles de desempeño Insuficiente en las dos competencias, son de género femenino, no presentan ningún tipo de discapacidad ni indicios de copia en la prueba, están ubicados en la región Caribe, se encuentran matriculados en Instituciones Educativas del sector Oficial, en la jornada de la mañana y calendario A, además el número instituciones en la región es inferior a 2000 y el número de estudiantes por región está entre 200.000 y 300.000; a diferencia del *cluster 0* estas son de la zona Rural, con entidad territorial certificada y las instituciones poseen un nivel socioeconómico que oscila entre 1 y 2.

- ✓ **Competencias Genéricas de Lenguaje y Competencias Ciudadanas (Lengcomp)**

De acuerdo a los resultados obtenidos en la figura 21, se generó 4 grupos similares en los cuatro niveles de desempeño contenidos en las competencias genéricas de Lenguaje y Competencias Ciudadanas, se realizó 4 iteraciones donde se clasifica el total de estudiantes (219748) en tres clusters. En el *cluster 0* a 89293 (41%) del total de estudiantes; en el *cluster 1* se agrupan 25008 (11%); en el *cluster 2* se encuentran 17071 (8%) y en el *cluster 3* se concentran 17071 (40%).

Al realizar la lectura de los resultados de cada *cluster*, se pueden obtener los siguientes patrones descriptivos:

***Cluster 0*:** El 41% de los estudiantes que presentaron la prueba de Lenguaje y Competencias Ciudadanas obtienen niveles de desempeño Mínimo en las dos competencias, son de género femenino, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, están ubicados en la región del eje cafetero, se encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Urbana, en la jornada de la mañana y calendario A, con entidad territorial no certificada, además el número instituciones en la región es superior a 4000 y

el número de estudiantes por región está entre 200.000 y 300.000 y las instituciones poseen un nivel socioeconómico que oscila entre 3 y 4.

***Cluster 1*:** El 11% de los estudiantes que presentaron la prueba de Lenguaje y Competencias Ciudadanas obtienen niveles de desempeño Mínimo en las dos competencias, son de género masculino, están ubicados en la región caribe, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, se encuentran matriculados en Instituciones Educativas del sector No Oficial de la zona Urbana, están en la jornada de la mañana y calendario A, con entidad territorial no certificada, además el número instituciones en la región es superior a 4000 y el número de estudiantes por región está entre 200.000 y 300.000 y las instituciones poseen un nivel socioeconómico que oscila entre 3 y 4.

***Cluster 2*:** El 8% de los estudiantes que presentaron la prueba de Lenguaje y Competencias Ciudadanas obtienen niveles de desempeño Mínimo en las dos competencias, son de género masculino, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, se encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Urbana con calendario A, con entidad territorial No certificada, y las instituciones poseen un nivel socioeconómico que oscila entre 3 y 4; a diferencia del *cluster1* los estudiantes están ubicados en la región de Bogotá, además el número instituciones en la región es inferior a 2000 pero el número de estudiantes por región sigue estando entre 200.000 y 300.000, están en jornada completa.

***Cluster 3*:** El 40% de los estudiantes que presentaron la prueba de Lenguaje y Competencias Ciudadanas obtienen niveles de desempeño Mínimo en las dos competencias , no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, se encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Urbana están en la jornada de la mañana y calendario A , además el número instituciones en la región es superior a 4000 y el

número de estudiantes por región está entre 200.000 y 300.000; a diferencia del *cluster 0* los estudiantes son de genero masculino, están ubicados en la región caribe, las instituciones poseen un nivel socioeconómico que oscila entre 1 y 2, con entidad territorial certificada.

- ✓ **Competencias Genéricas de Lenguaje y Matemáticas (Lengmate)**

De acuerdo a los resultados obtenidos en la figura 23, se generó 3 grupos similares en los cuatro niveles de desempeño contenidos en las competencias genéricas de Lenguaje y Matemáticas se realizó 3 iteraciones donde se clasifica el total de estudiantes (671052) en tres clusters. En el *cluster 0* a 136263 (20%) del total de estudiantes; en el *cluster 1* se agrupan 209072 (31%) y en el *cluster 2* se encuentran 325717 (49%).

Al realizar la lectura de los resultados de cada *cluster*, se pueden obtener los siguientes patrones descriptivos:

***Cluster 0*:** El 20% de los estudiantes que presentaron la prueba de Lenguaje y Matemáticas obtienen niveles de desempeño Insuficiente en las dos competencias, son de género Masculino, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, están ubicados en la región centro oriente, se encuentran matriculados en Instituciones Educativas del sector No Oficial de la zona Urbana, en la jornada Completa y calendario A, con entidad territorial no certificada, además el número instituciones en la región está entre 2000 y 4000 y el número de estudiantes por región está entre 200.000 y 300.000 y las instituciones poseen un nivel socioeconómico que oscila entre 3 y 4.

***Cluster 1*:** El 31% de los estudiantes que presentaron la prueba de Lenguaje y Matemáticas obtienen niveles de desempeño Insuficiente en las dos competencias, son de género Femenino, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, y son de calendario A; a diferencia del *cluster 0* están ubicados en la región Caribe, se encuentran

matriculados en Instituciones Educativas del sector Oficial de la zona Rural, en la jornada de la Mañana, con entidad territorial certificada, además el número instituciones en la región es superior a 4000 aunque el número de estudiantes por región sigue estando entre 200.000 y 300.000 y las instituciones poseen un nivel socioeconómico que oscila entre 1 y 2.

***Cluster 2*:** El 49% de los estudiantes que presentaron la prueba de Lenguaje y Matemáticas obtienen niveles de desempeño Mínimo en las dos competencias, son de género Masculino, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, en la jornada de la Mañana y calendario A, con entidad territorial no certificada, además el número instituciones en la región está entre 2000 y 4000 y el número de estudiantes por región está entre 200.000 y 300.000 y las instituciones poseen un nivel socioeconómico que oscila entre 3 y 4. A diferencia del *cluster 0* están ubicados en la región centro oriente, se encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Urbana, además el número instituciones en la región es superior a 4000 aunque el número de estudiantes por región sigue estando entre 200.000 y 300.000.

- ✓ **Competencias genéricas de Matemáticas y Ciencias Naturales (Matecien)**

De acuerdo a los resultados obtenidos en la figura 25, se generó 3 grupos similares en los cuatro niveles de desempeño contenidos en las competencias genéricas de Matemáticas Ciencias Naturales se realizó 4 iteraciones donde se clasifica el total de estudiantes (444318) en tres clusters. En el *cluster 0* a 235607 (53%) del total de estudiantes; en el *cluster 1* se agrupan 118793 (27%) y en el *cluster 2* se encuentran 89918 (20%).

Al realizar la lectura de los resultados de cada *cluster*, se pueden obtener los siguientes patrones descriptivos:

***Cluster 0*:** El 53% de los estudiantes que presentaron la prueba de Matemáticas y Ciencias Naturales obtienen niveles de desempeño Insuficiente en las dos competencias, son de género

Femenino, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, están ubicados en la región Caribe, se encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Urbana, en la jornada de la Mañana y calendario A, con entidad territorial no certificada, además el número instituciones en la región es superior a 4000 y el número de estudiantes por región está entre 200.000 y 300.000 y las instituciones poseen un nivel socioeconómico que oscila entre 3 y 4.

***Cluster 1*:** El 27% de los estudiantes que presentaron la prueba de Matemáticas y Ciencias Naturales obtienen niveles de desempeño Mínimo en las dos competencias, son de género Masculino, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, están ubicados en la región Caribe, se encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Urbana, en la jornada de la Mañana y calendario A, con entidad territorial no certificada, además el número instituciones en la región es superior a 4000 y el número de estudiantes por región está entre 200.000 y 300.000, pero a diferencia del *cluster 0* las instituciones poseen un nivel socioeconómico que oscila entre 1 y 2.

***Cluster 2*:** El 20% de los estudiantes que presentaron la prueba de Matemáticas y Ciencias Naturales obtienen niveles de desempeño Insuficiente en las dos competencias, son de género Masculino, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, están ubicados en la región Caribe, en la jornada de la Mañana y calendario A, además el número instituciones en la región es superior a 4000 y el número de estudiantes por región está entre 200.000 y 300.000, a diferencia del *cluster 0* se encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Rural, con entidad territorial certificada y las instituciones poseen un nivel socioeconómico que oscila entre 1 y 2.

✓

- ✓ **Competencias Genéricas de Matemáticas y Competencias Ciudadanas (Matecomp)**

De acuerdo a los resultados obtenidos en la figura 27, se generó 3 grupos similares en los cuatro niveles de desempeño contenidos en las competencias genéricas de Matemáticas y Competencias Ciudadanas, se realizó 4 iteraciones donde se clasifica el total de estudiantes (220248) en tres clusters. En el *cluster 0* a 112734 (51%) del total de estudiantes; en el *cluster 1* se agrupan 50779 (23%) y en el *cluster 2* se encuentran 56735 (26%).

Al realizar la lectura de los resultados de cada *cluster*, se pueden obtener los siguientes patrones descriptivos:

***Cluster 0*:** El 51% de los estudiantes que presentaron la prueba de Matemáticas y Competencias Ciudadanas obtienen niveles de desempeño Mínimo en las dos competencias, son de género Masculino, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, están ubicados en la región Caribe, se encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Urbana, en la jornada de la Mañana y calendario A, con entidad territorial certificada, además el número instituciones en la región es superior a 4000 y el número de estudiantes por región está entre 200.000 y 300.000 y las instituciones poseen un nivel socioeconómico que oscila entre 3 y 4.

***Cluster 1*:** El 23% de los estudiantes que presentaron la prueba de Matemáticas y Competencias Ciudadanas obtienen niveles de desempeño Mínimo en las dos competencias, son de género Femenino, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, están ubicados en la región Caribe, en la jornada de la Mañana y calendario A, con entidad territorial certificada, además el número instituciones en la región es superior a 4000 y el número de estudiantes por región está entre 200.000 y 300.000 a diferencia del *cluster 0* se

encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Rural y las instituciones poseen un nivel socioeconómico que oscila entre 1 y 2.

***Cluster 2*:** El 26% de los estudiantes que presentaron la prueba de Matemáticas y Competencias Ciudadanas obtienen niveles de desempeño Mínimo en las dos competencias, son de género Femenino, no presentan ningún tipo de discapacidad cognitiva ni indicios de copia en la prueba, con calendario A, están ubicados en la región Caribe, a diferencia del *cluster 1* se encuentran matriculados en Instituciones Educativas del sector Oficial de la zona Urbana, en la jornada de la Tarde, con entidad territorial No Certificada, además el número instituciones en la región es inferior a 2000 aunque el número de estudiantes por región permanece entre 200.000 y 300.000 y las instituciones poseen un nivel socioeconómico que oscila entre 3 y 4.

Los resultados obtenidos en los clusters predomina el género masculino y las regiones del Caribe, Bogotá y Eje Cafetero; en el tipo de establecimiento predomina las instituciones del sector oficial ubicadas en zona urbana. En cuanto a los niveles de desempeño se destaca el nivel insuficiente en la prueba de Lenguaje y Ciencias Naturales, Lenguaje y Matemáticas, Matemáticas y Ciencias Naturales; en el nivel de desempeño mínimo está la prueba de Lenguaje y Competencias Ciudadanas y Matemáticas y Competencias Ciudadanas.

El modelo descriptivo obtenido, a través de la técnica de clustering con el algoritmo k-means ha ofrecido resultados que nos ha permitido conocer los factores que inciden a un determinado nivel de desempeño. Sin embargo a partir de las investigaciones realizadas descritas en el capítulo 2, hubiese sido importante contar con información adicional en los aspectos socioeconómicos del estudiante ya que la base de datos proporcionada por el ICFES es limitada en estos aspectos.

3.6 Implementación

En esta fase, el conocimiento descubierto se incorpora al existente y podrá ser utilizado en los procesos de toma de decisiones de las instituciones gubernamentales y académicas que velan por la calidad de la educación de básica primaria en Colombia. Una vez sean intervenidas las instituciones educativas que presentan factores asociados al rendimiento escolar como son: zona urbana, nivel socioeconómico bajo, pertenecen al sector oficial, tienen calendario A y que los resultados en el desempeño académico son mínimos en las Pruebas Saber 5, será posible analizar los resultados y determinar sus efectos.

4 Discusión

El objetivo de esta investigación fue descubrir factores asociados al desempeño académico de las pruebas Saber 5 con técnicas descriptivas de minería de datos, a partir de la información almacenada en el repositorio de datos del ICFES en el periodo 2014-2016. Para cumplir este objetivo se escogieron las tareas de asociación y clustering. Se analizaron 2’038.120 casos de estudiantes que presentaron dicha prueba. Se obtuvieron diferentes tipos de patrones de desempeño dependiendo de la técnica de minería de datos aplicada. Los patrones resultantes de las reglas de asociación descubiertas, determinan cuales factores o atributos, que inciden en el desempeño académico, se están presentando juntos, y los clusters obtenidos, permiten conocer las características de los estudiantes que forman los distintos grupos, basado en sus similitudes.

Los modelos de asociación, basados en el algoritmo Apriori, construidos para cada par de competencias genéricas, al igual que los modelos de agrupación, generados con el algoritmo simple k-means, muestran que los factores como el género, jornada, calendario, nivel socioeconómico, zona del establecimiento, tipo de establecimiento, número de estudiantes por zona y número de instituciones por zona se asocian al nivel de desempeño mínimo. Este resultado está en línea con el estudio hecho por la OCDE (2016), en donde se destaca el esfuerzo del gobierno nacional por integrar programas que apoyen los procesos de mejoramiento de la calidad educativa, pese a esto los estudiantes continúan obteniendo competencias bajas en relación con otros países. Los resultados de la investigación indican que la competencia genérica en lenguaje sigue siendo baja comparada con el promedio de la OCDE y con otros países latinoamericanos. El desempeño en la competencia genérica de matemáticas y ciencias no ha cambiado. En matemáticas, los estudiantes colombianos de 15 años están, en promedio, atrasados más de tres años (118 puntos) con respecto a sus pares de países miembros de la OCDE (OCDE, 2014b). El Tercer Estudio

Regional Comparativo y Explicativo (TERCE) el cual evalúa las competencias en matemáticas, ciencias y escritura de los estudiantes latinoamericanos de los Grados 3 y 6, muestra que los estudiantes colombianos empiezan a atrasarse con respecto a sus países vecinos como Chile, Costa Rica y México, en los primeros años de educación (Oficina de la UNESCO de Santiago, 2015). En pruebas internacionales como la PISA los resultados muestran que Colombia se encuentra en los niveles bajos de desempeño; pero también evalúa el efecto que tienen sobre el aprendizaje las variables socioeconómicas y culturales, así como las características del sistema escolar y las instituciones educativas.

Además se encontraron factores que están estrechamente relacionados con el establecimiento educativo, como: el sector, la zona, la jornada y el nivel socioeconómico. En otros estudios se han analizado factores asociados al desempeño académico y coinciden con algunos de los que se han descubierto en esta investigación, como el realizado por Gaviria y Barrientos (2001a y b), quienes determinaron que las características asociadas al plantel educativo inciden de manera significativa en el rendimiento y lo hacen en mayor medida que las variables socioeconómicas. Adicionalmente, evidencian que existe una brecha pronunciada entre los resultados para instituciones oficiales y privadas: estas últimas alcanzan mayores logros en las pruebas. Aunque otros autores como Coleman (1966) en su informe sobre rendimiento académico no está de acuerdo al concluir que el rendimiento escolar este influenciado en gran medida por las características socioeconómicas de los estudiantes y que las variables asociadas a la institución educativa tienen poco o ningún efecto sobre las diferencias en el desempeño escolar. En contraposición con la teoría de Coleman y lo estudiado en la revisión de literatura encontrada sobre estos factores, se confirma que las variables institucionales, juegan un papel preponderante en el desempeño, puesto que abarcan otros aspectos como: la infraestructura física, lo espacios comunes, biblioteca,

programas académicos e institucionales por mencionar algunos, que según Zapata, A., et al (2015) en su estudio sobre factores institucionales en el rendimiento académico estos son incidentes en los resultados académicos. Cabe mencionar que el rendimiento académico lo asocian algunos autores como Murillo (2007) con la eficacia escolar, y ha sido una línea de investigación ampliamente estudiada con el fin de aportar al mejoramiento de la calidad educativa. La cual es entendida como el proceso de aprendizaje dependiente de diversos factores como son aspectos económicos, sociales, políticos, familiares, cognitivos, entre otros (Erazo, O. 2012).

De otra parte, en Colombia se han realizado varios estudios similares con el fin de determinar los factores que influyen en el rendimiento académico de los estudiantes. Se destaca el realizado por Gaviria y Barrientos (2001a y b). Allí, los autores analizan los resultados de las pruebas de Estado. Ellos encontraron que las características asociadas al plantel educativo inciden de manera significativa en el rendimiento y lo hacen en mayor medida que las variables socioeconómicas. Sin embargo, no desconocen que el nivel de educación de los padres juega un papel fundamental en el desempeño. Adicionalmente, evidencian que existe una brecha pronunciada entre los resultados para instituciones oficiales y privadas: estas últimas alcanzan mayores logros en las pruebas. Todo lo anterior deja entredicho la aplicación de la hipótesis de Coleman (1966) para Colombia.

Analizando uno de los patrones generados por los modelos, como lo es el sector del establecimiento educativo (oficial y no oficial), estudios muestran que las características asociadas al plantel educativo inciden de manera significativa en el rendimiento y lo hacen en mayor medida que las variables socioeconómicas. Esto debido a que existen concepciones como las expuestas en Castro, G., et al (2016) en donde se relaciona el tipo de colegio con las características de los profesores, los recursos didácticos, el número de estudiantes, la infraestructura educativa y los

niveles de gasto en educación. Por ejemplo, Piñeros y Rodríguez (1999) demuestran la importancia de estos aspectos en el aprendizaje, de hecho, las diferencias entre las escuelas privadas y las públicas repercuten en los resultados de las pruebas de estado. Concluyen que hay un mejor desempeño de los estudiantes, cuando se asisten a un colegio privado.

Por otra parte, otro de los patrones generados es la zona del establecimiento educativo (Urbana o Rural), esta variable está estrechamente ligada al nivel socioeconómico del establecimiento, puesto que, a mejores condiciones en infraestructura, accesibilidad a recursos y materiales presentan mejores desempeños académicos. Aunque la zona rural está creciendo y desarrollándose más rápido que la zona urbana gracias a la implementación de los programas de acción para el fortalecimiento educativo en el campo como el PER (Proyecto de Educación Rural), aún hay mucho por hacer por la reducción de la brecha educativa, ya que según los resultados generados por las reglas de asociación y agrupación, los establecimientos educativos de la zona urbana presenta mejores desempeños que la zona rural. Blackwell y McLaughlin (1999), si encuentran que la variable de localización rural/urbana es significativa a la hora de explicar el rendimiento. Esto radica en que las características de los estudiantes, sus familias y las escuelas son diferentes en estos dos grupos. Los estudiantes de las zonas rurales suelen formar parte de familias con pocos recursos económicos, sus padres tienen bajo nivel de educación y las escuelas a las que asisten cuentan con escasas dotaciones y generalmente son más pequeñas que las escuelas urbanas.

Como último factor, el género del estudiante, los resultados indican que el género masculino obtiene mejores desempeños en la competencia genérica de Matemáticas, y para el género femenino en la competencia de Lenguaje. La misma diferencia se muestra en el estudio de la UNESCO (2010), tras un análisis de varias investigaciones, entre ellas la realizada por el

SIMCE[4] sobre la asociación entre el rendimiento de los estudiantes y el género. El resultado es que en la prueba del año 2004 de 8° Básico, las mujeres superan a los hombres en Lenguaje y Comunicación, los hombres alcanzan mayores niveles de aprendizaje que las mujeres en Educación Matemática. Así como en el SIMCE, en PISA 2001 las mujeres también tuvieron un mayor logro que los hombres en Lenguaje. En TIMSS 2002, los hombres alcanzan un mayor nivel de aprendizaje que las mujeres, en Matemática y Ciencias.

En resumen, entre los patrones de desempeño académico de las competencias genéricas de las pruebas saber 5, descubiertos y considerados en esta discusión, los factores asociados relacionados con los planteles educativos se encuentran relacionados con atributos como el sector del establecimiento (oficial y no oficial), la zona donde se encuentra ubicado, rural o urbano, el nivel socioeconómico, y en relación al estudiante, el género. Entre otros atributos se destacan, el área del conocimiento, el número de instituciones en la zona, el número de estudiantes por zona; atributos que, en relación con el desempeño académico, de alguna manera, son considerados en los resultados de investigaciones de diversos autores.

Por otra parte, también se encontró que existe asociación para el buen desempeño académico en la competencia genérica de competencias ciudadanas con respecto al número de estudiantes por zona y al número de instituciones por zona. De los resultados se deduce que a mayor cantidad de estudiantes mejores son los desempeños en la competencia mencionada.

[4] SIMCE, 2006: Análisis de las diferencias de logro en el aprendizaje escolar entre hombres y mujeres.

5 Conclusiones

El objetivo de esta investigación fue descubrir factores asociados al desempeño académico de los estudiantes de las Instituciones Educativas Colombianas a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES, utilizando técnicas descriptivas de Minería de Datos. Dicho objetivo se logró en su totalidad.

Agrupando los resultados obtenidos a través de las técnicas de minería de datos, para las competencias genéricas de Lenguaje, Ciencias Naturales, Matemáticas y Competencias Ciudadanas, que son las que se evalúan en la Prueba Saber 5, se obtuvo que los factores descubiertos están asociados al desempeño mínimo., donde toma mayor relevancia la zona urbana en la que se ubican los establecimientos educativos, el sector oficial, el calendario A, el tipo de entidad territorial no certificada, la jornada de la mañana y un número alto de instituciones por zona. Además, se destaca el género en relación a los resultados obtenidos en ambas competencias.

Los resultados obtenidos a través de las tareas de asociación y clustering generan modelos consistentes con la realidad observada basándose en los datos que se encuentran almacenados en el repositorio de datos del ICFES.

Según los resultados obtenidos en la fase de evaluación, se puede concluir que es necesaria la inclusión de factores socioeconómicos del estudiante, ya que su identificación (NUIT, Tarjeta de Identidad o Registro Civil) no se evidencia en las bases de datos y por tanto no es posible localizar esta información en otras fuentes de información como el DANE, dicha información es de vital relevancia para el descubrimiento de otros factores que inciden en el desempeño académico

de las Pruebas Saber 5, entre estos están las condiciones económicas que afectan en gran medida la probabilidad de no ubicarse en el nivel de desempeño insuficiente o mínimo para las cuatro competencias genéricas. Lo anterior se puede explicar en el sentido en que un entorno socioeconómico propicio le permite al estudiante contar con las condiciones esenciales para dedicarse sin inconvenientes a sus estudios. Ya que el niño o niña puede acceder a una mejor alimentación, transporte, infraestructura, recursos tecnológicos, entre otros aspectos.

Es necesario reconocer que los resultados de los modelos de asociación con el algoritmo Apriori, y de los modelos de agrupación con el algoritmo k-means, obtenidos para las cuatro competencias genéricas, son satisfactorios para poder obtener los patrones que asocian los factores socioeconómicos, académicos e institucionales con el desempeño académico de los estudiantes que presentaron las pruebas Saber 5 en el periodo 2014-2016.

La evaluación, análisis y utilidad de los factores descubiertos permitirá incorporar el conocimiento descubierto al existente y se integrará a los procesos de toma de decisiones de las instituciones gubernamentales y educativas que velan por la calidad de la educación en la República de Colombia. Una vez se realicen los debidos procesos de intervención en las instituciones educativas que presentaron factores que inciden en el rendimiento académico de las Pruebas Saber 5, será posible analizar los resultados y determinar sus efectos.

6 Recomendaciones

Haciendo un contraste entre las variables de las bases de datos del ICFES y el marco de factores asociados, se observó que existe un número limitado de atributos para realizar el análisis a mayor profundidad. Se sugiere ampliar la información socioeconómica del estudiante ya que está puede ser determinante para el desempeño académico.

Se recomienda continuar investigando sobre los factores que inciden en el rendimiento académico de las pruebas saber, seguramente identificando nuevos atributos o revaluando los utilizados, mejorando las técnicas de limpieza y transformación utilizadas o usando otras técnicas de minería de datos que nos permitan mejores resultados.

Para futuros trabajos de investigación se debería incorporar otras variables que forman parte del modelo CIPP (Contexto, Insumos, Procesos y Productos), las mismas que no pudieron ser tomadas en cuenta puesto que las bases de datos proporcionadas por el ICFES no contaban con dicha información. Algunos de los aspectos que se tendrían que incluir son: edad, nivel socioeconómico familiar, violencia en el entorno del hogar, involucramiento parental, violencia en el entorno de la institución educativa, antecedentes escolares, entre otras.

7. Referencias Bibliográficas

Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A.I. (1996): Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, AAAI Press /The MIT Press.

Agrawal, R., Imielinski, T. and Swami, A.N. (1993) Mining Association Rules between Sets of Items in Large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 22, 207-216.Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. VLDB'94 487-499, Santiago, Chile.

Ajani, A., & Akinyele, R. (2014). Effects of Student-Teacher Ratio on Academic Achievement of Selected Secondary School Students in Port Harcourt Metropolis, Nigeria. Journal of Education and Practice, 5(24), 100-106.

Ball, S. (2012). Privatizaçoa da educaçao e novas subjetividades: contornos e esdobramientos das políticas (pós) neoliberais. Revista Brasileira de Educação, 18(53), 457-466.

Bowen, N. K., & Bowen, G. L. (1999). Effects of Crime and Violence in Neighborhoods and Schools on the School Behavior and Performance of Adolescents. Journal of Adolescent Research, 14(3), 319-342. doi:10.1177/0743558499143003.

Castro, G., Díaz, M. y Tobar, J. (2016). Causas de las diferencias en desempeño escolar entre los colegios públicos y privados: Colombia en las pruebas saber11 2014. Documentos de trabajo FCEA ISSN 1909-4469 / ISSNe 2422-4642. Año 2016 N°26.

Celis, M. T., Jiménez, Ó. A., & Jaramillo, J. F. (2012). ¿Cuál es la brecha de la calidad educativa en Colombia en la educación media y en la superior? En Estudios sobre calidad de la educación en Colombia (pp. 67-98). Bogotá: Instituto Colombiano para la Evaluación de la Educación ICFES.

Chapman, P., Clinton, J., Randy, K., Khabaza, T., Reinartz, T., Shearer, C. y Wirth, R. (2000). CRISP-DM 1.0 Step-by-Step Data Mining Guide. Recuperado de http://www.crisp-dm.org/ CRISPWP-0800.pdf.

Chaux, E. (2009). Citizenship Competencies in the Midst of a Violent Political Conflict: The Colombian Educational Response. Harvard Educational Review, 79(1), 84–93. https://doi.org/10.17763/haer.79.1.d2566q027573h219.

Chen, M., Han, J. y Yu, P. (1996). Data Mining: An Overview from Database Perspective. ieee Transactions on *Knowledge and Data Engineering*.

Chica, S., Galvis, D., Ramírez, A. (2010) Determinantes del rendimiento académico en Colombia: pruebas ICFES Saber 11, 2009. Revista Universidad EAFIT, Vol. 46, Núm. 160. ISSN: 0120341X. Medellín, Colombia.

Chubb, J. (2001). The profit motive. The private can the public. Education Next, 1(1).

Cohen, J. (1988). Análisis de poder estadístico para las Ciencias del comportamiento (2da. ed.). Nueva Jersey: Lawrence Erlbaum.

Cohen, J., McCabe, L., Michelli, N. M., & Pickeral, T. (2009). School climate: Research, policy, practice, and teacher education. *Teachers College Record*, 111(1), 180-213.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F. and York, R., (1966). Equality of Educational Opportunity. National Center for Educational Statistics. Report number OE38001. US Government Printing Office. Washington, USA.

Crenshaw, M. (2003). The relationships among school size, school climate variables, and achievement ratings in South Carolina high schools: A conceptual model. University of South Carolina, Columbia.

De Moya, A. & Rodríguez, R., (2003) La contribución de las reglas de asociación a la Minería de Datos. Revista Tecnura ll semestre, 94-109.

Duflo, E., Dupas, P. & Kremer, M. (2011). Peer effects, Teacher Incentives and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. American Economic Review. 101(5), pp. 1739-1774.

Elmore, R. (2010). Mejorando la escuela desde la sala de clases. Área de Educación Fundación Chile.

Erazo, O. (2012). El rendimiento académico, un fenómeno de múltiples relaciones y complejidades. Revista Vanguardia Psicológica / Año 2 / Volumen 2 / Número 2, octubre-marzo / pp. 144-173 / ISSN 2216-0701.

Espínola, V. (2005). Educación para la ciudadanía y la democracia para un mundo globalizado: una perspectiva comparativa. Washington: Banco Interamericano de Desarrollo.

Fernández, H. (2005). Como interpretar la evaluación Pruebas Saber. Subdirección de Estándares y Evaluación. Ministerio de Educación Nacional. Bogotá, Colombia.

Ferrer, G. (2006). Sistemas de evaluación de aprendizajes en América Latina: balance y desafíos: PREAL.

García, D. (s.f.). Manual de Weka. Recuperado a partir de http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf

Goldin, C., Katz, L., & Kuziemko, I. (2006). The homecoming of american college women: the reversal of the college gender gap.

Gonzáles, C., Caso, J., Díaz, K. y López, M. (2012). Rendimiento académico y factores asociados. Aportaciones de algunas evaluaciones a gran escala. Bordón 64(2), 2012, 51-68. ISSN: 0210-5934.

Guerra, L. (2008). Primeros pasos con Knime. Recuperado a partir de http://laurel.datsi.fi.upm.es/_media/docencia/cursos/inap/ejemplodm.pdf

Gutiérrez, Y. (2015). Relación entre la estructura familiar y el rendimiento académico en el área de matemáticas. Instituto Latinoamericano de altos estudios –ILAE. Editorial Milla. Bogotá, Colombia.

Han, J. y Kamber, M. (2001). Data Mining Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.

Hernández, J., Ramírez, M. & Ferri, C. (2004). Introducción a la Minería de Datos. Editorial Pearson Educación. S.A., Madrid. ISBN: 978-84-205-4091-7.

Hernández, J., Ramírez, M. & Ferri, C. (2005). Introducción a la Minería de Datos. Madrid: Pearson Prentice Hall.

Holmes, C. (1989). Grade Level Retention Effects: A Meta-Analysis of Research Studies in L.A Shepard and M. L. Smith. (Eds), Flunking Grades: Research and Policies on Retention. The Falmer Press.

Huang, Z. (1997). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. En: Research Issues on Data Mining and Knowledge Discovery.

Huang, Z. (1998). Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery.

ICFES (2011). Informe técnico Saber 5° y 9° 2009. Instituto Colombiano para la Evaluación de la Educación (ICFES). ISBN de la versión electrónica: 978-958-11-0578-6. Bogotá, Colombia.

ICFES (2011). Lineamientos generales Saber 2009 5° y 9°. Instituto Colombiano para la Evaluación de la Educación (ICFES). ISBN de la versión electrónica: 978-958-11-0489-5. Bogotá, marzo de 2009.

ICFES (2011). Marco de factores Asociados Saber 3°, 5° y 9° 2016. Instituto Colombiano para la Evaluación de la Educación (ICFES). ISBN de la versión digital: En trámite. Bogotá, Colombia.

ICFES (2011). SABER 5° y 9° 2009 Síntesis de resultados de factores asociados. Instituto Colombiano para la Evaluación de la Educación (ICFES). ISBN de la versión electrónica: 978-958-11-0574-8. Bogotá, Colombia.

ICFES (2012). Estudios sobre calidad de la educación en Colombia. Bogotá, D.C., noviembre de 2012. ISBN de la versión electrónica: 978-958-11-0595-3

ICFES (2013). Alineación del examen SABER 11°. Sistema Nacional de Evaluación Estandarizada de la Educación, Instituto Colombiano para la Evaluación de la Educación (ICFES). Bogotá, Colombia.

ICFES (2014). Alineación del examen SABER 11° Lineamientos generales 2014-2 Sistema Nacional de Evaluación Estandarizada de la Educación, Instituto Colombiano para la Evaluación de la Educación (ICFES). ISBN: 9789581106301. Bogotá, Colombia.

Jafet, C. & Martínez, C. (2016). Factores del contexto socio-cultural y familiar que influyen en el bajo rendimiento académico de los estudiantes de grado 601 de la Institución Educativa General Santander de la Sede Principal J.T. del municipio de Soacha. Bogotá: Corporación Universitaria Minuto de Dios.

Jiménez, A. & Álvarez, H. (2010). Minería de Datos en la Educación. Universidad Carlos III de Madrid Avda. De la Universidad, 3028911, Leganés (Madrid-España).

Jiménez, L., & Rengifo, P. (Diciembre de 2010). Al interior de una máquina de soporte vectorial. (U. d. Valle, Ed.) REVISTA DE CIENCIA, 14, 73-85.

Kurtz-Dostes y Scheneider, (1994) self-concept attributional beliefs and school achievement: a longitudinal analysis. Contemporary educational psichology 19(2) 199-216.

Lamdin, D. (1995). Sting for effect of school size on student achievement within a school district. Education Economics, 3, 33-42.

Leithwood, K. (1994). Leadership for school restructuring. Educational administration quarterly, 30(4), 498-518.

Lester, K; Garofalo, J; & Kroll, D (1989). Self-Confidence, Interest, Beliefs, and Metacognition: Key Influences on Problem-Solving Behavior. In D.B. McLeod & V. M. Adams (Eds.), Affect and mathematical problem solving (75-88). Springer New York.

Martín, S. (2015). Pruebas Saber de lenguaje 3° y 5°: Posibilidades y retos desde la perspectiva de la evaluación formativa. Universidad Pedagógica Nacional. Bogotá, Colombia.

MEN (2008). Colombia: qué y cómo mejorar a partir de la prueba PISA. Altablero n° 44.Enero-Marzo 2008.Recuperado de http://www.mineducacion.gov.co/1621/article-162392.html.

Mendoza, C., y Bustamante, C. (2013). Estudios de correlación. Revista De Actualización Clínica, 33, 1690-1694.

Moody, J. & Darken, C. (1989) Fast Learning in networks of locally tuned processing units. Neural Computation.

MURILLO, F.J. (Coord.) (2005). Estudios sobre Eficacia Escolar en Iberoamérica. 15 buenas investigaciones. Bogotá: Convenio Andrés Bello.

Murillo, J. (2007). Investigación iberoamericana sobre eficacia escolar. Bogotá: Convenio Andrés Bello.

Niederle, M., & Vesterlund, L. (2010). Explaining the Gender Gap in Math Test Scores: he Role of Competition. The Journal of Economic Perspectives, 24(2), 129-144. http://web.stanford. edu/~niederle/NV.JEP.pdf

OCDE (2016). Education in Colombia. Versión en inglés: (ISBN 9789264250598/http://dx.doi.org/10.1787/9789264250604-en). Recuperado de http://www.oecd.org/edu/school/Education-in-Colombia-Highlights.pdf

OCDE. (2003). Learners for life. Student approaches to learning. Results from PISA 200. Paris: OECD Publishing.

OCDE. (2010b). Pisa 2009 Results: Learning to Learn - Student Engagement, Strategies and Practices.

OECD. (2012). Starting Strong III: a quality toolbox for early childhood education and care. Paris: OECD Publishing.

OECD. (2015). Students, Computers and Learning: Making the Connection. PISA. OECD Publishing. http://dx.doi.org/10.1787/9789264239555-en.

OECD. (2016a). PISA 2015 Assessment and Analytical Framework: OECD Publishing.

Orea, S. V., Vargas, A. S., & Alonso, M. G. (2005). Minería de Datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. Ene, 779(73), 33.

Ospina J. (2006) La motivación, motor del aprendizaje, Rev. Cienc. Salud. Bogotá (Colombia) 4 (Especial): 158-160.

Piatetsky-Shapiro, G., Brachman, R. y Khabaza, T. (1996). An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications. Association for the Advancement of Artificial Intelligence [aaai], mit Press. Recuperado de http://www.aaai.org/Papers/KDD/1996/KDD96-015.pdf.

Piñeros, L. & Rodríguez, A., 1999. School inputs in secondary education and their effects on academic achievement: a study in Colombia, s.l.: LCSHD Paper Series No. 36. Human Development Department, World Bank.

Pomerantz, E. M., Moorman, E. A., & Litwack, S. D. (2007). The How, Whom, and Why of Parents' Involvement in Children's Academic Lives: More Is Not Always Better. Review of Educational Research, 77(3), 373-410. doi:10.3102/003465430305567.

Reimers, F., DeShano da Silva, C., & Treviño, E. (2006). Where is the Education in Conditional Cash Transfer in Education? UNESCO Institute for Statistics. Retrieved from http://www.uis. unesco.org/Library/Documents/WP1-06-en.pdf.

Reyes, J. & García, R., 2005. El proceso de descubrimiento de conocimiento en bases de datos. Revista Ingenierías, Enero-Marzo 2005, Vol. Vlll, No 26, 37-47.

Riquelme, J. Ruiz, R. & Gilbert, K., (2006) Minería de datos: Conceptos y tendencias Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. No.29 (2006), pp. 11-18.

Roderick, M. (1994). Grade Retention and School Dropout: Investigating the Association. American Educational Research Journal. Vol. 31(4). Pp. 729-759.

Rodríguez, H. (2007). El paradigma de las competencias hacia la educación superior. Revista Facultad de Ciencias Económicas. Universidad Militar Nueva Granada. V. XV, N. 1, 145165.

Rodriguez, E. (2014). La influencia de los factores familiares en el rendimiento académico. Universidad de Valladolid.

Ryan, R. M; & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions; University of Rochester.

Rychen, D.S. y Salganik, L.H. (eds.)Traducción al español: Las competencias clave para el bienestar personal, económico y social (1ª ed. en español, 2006).

Stevenson, K. (1996). Elementary school capacity: What size is the right size? CEFP Journal, 33(4), 10-14.

Timarán, R., Calderón, A. y Jiménez, J. (2013a). Aplicación de la Minería de Datos en la extracción de perfiles de deserción estudiantil. Ventana Informática, No 28, 2538.

Timarán, R., Calderón, A. y Jiménez, J. (2013b). La Minería de Datos como un método innovador para la detección de patrones de deserción estudiantil en programas de pregrado en Instituciones de Educación Superior. En Memorias Foro Mundial de Educación en Ingeniería, WEEF 2013. Cartagena, Colombia: ACOFI & IFEES.

Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, J. J., Hidalgo-Troya, A. y Alvarado- Pérez, J. C. (2016). Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. Bogotá: Ediciones Universidad Cooperativa de Colombia.

Torrado, M. (2000). Educar para el desarrollo de las competencias: Una propuesta para reflexionar. En Competencias y proyecto pedagógico. Universidad Nacional.

Torres, J., Pachajoa, L. y Pantoja, R. (2014). Resultados de las Pruebas Saber en el grado quinto del área de las ciencias naturales en tres instituciones educativas oficiales del municipio de Pasto. Revista Fedumar Pedagogía y Educación, 1(1), 55-69.

Unesco. (2005). Guidelines for inclusion: ensuring access to education for all. Paris: United Nations Educational, Scientific and Cultural Organization.

Unesco. (2010). Factores asociados al logro cognitivo de los estudiantes en América Latina y el Caribe. Santiago: Oficina Regional de Educación de la UNESCO para América Latina y el Caribe.

Unesco. (2015a) ¿Es la repitencia efectiva? (Vol. 1). Santiago: Oficina Regional de Educación para América Latina y el Caribe.

Unesco. (2015b). Factores Asociados. Informe de resultados Tercer Estudio Regional Comparativo y Explicativo. Santiago: Oficina Regional de Educación para América Latina y el Caribe (OREALC/Unesco Santiago). Retrieved from Santiago: http://unesdoc.unesco.org/ images/0024/002435/243533s.pdf

Unesco. (2015b). Factores Asociados. Informe de resultados Tercer Estudio Regional Comparativo y Explicativo. Santiago: Oficina Regional de Educación para América Latina y el Caribe (OREALC/Unesco Santiago). Retrieved from Santiago: http://unesdoc.unesco.org/images/0024/002435/243533s.pdf

Unesco. (2016a). Recomendaciones de políticas educativas para América Latina en base al TERCE. Santiago: Unesco.

UNESCO-UIS (2015), “Browse by theme: Education”, Data Centre, UNESCO Institute for Statistics, www.uis.unesco.org/DataCentre/Pages/BrowseEducation.aspx.

Valero, S. (2009). Aplicación de técnicas de Minería de Datos para predecir deserción. Puebla, México: Universidad Tecnológica de Izúcar de Matamoros. Recuperado de http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf.

Valero, S., Salvador, A. y García, M. (2010). Minería de Datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. Puebla, México: Universidad Tecnológica de Izúcar de Matamoros. Recuperado de www.utim.edu.mx/~svalero/docs/e1.pdf.

Vila, M., Sanchez, D., & Cerda, L. (2004). Reglas de asociación aplicadas a la detección de fraude con tarjeta de crédito. Actas del XII Congreso Español sobre Tecnologías Lógica Fuzzy, (págs. 15-17).

Weinstein, C. E., Husman, J., & Dierking, D. R. (2000). Self-regulation interventions with a focus on learning strategies. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), Handbook of selfregulation (pp. 727–747). San Diego, CA: Academic Press.

Westheimer, J., & Kahne, J. (2004). What kind of citizen? The politics of educating for democracy. American Educational Research Journal, 41(2), 237-269.